

ENCYCLOPEDIA OF
MEASUREMENT
AND
STATISTICS



EDITED BY
NEIL J. SALKIND

ENCYCLOPEDIA OF
MEASUREMENT

AND
STATISTICS

*This encyclopedia is dedicated to the memory of Rick Snyder—friend,
colleague, and leading scholar in the field of positive psychology.
Rick always gave far more than he took and made this world a better place.*

ENCYCLOPEDIA OF
MEASUREMENT
AND
STATISTICS

VOLUME **1**

EDITOR
NEIL J. SALKIND
UNIVERSITY OF KANSAS

MANAGING EDITOR
KRISTIN RASMUSSEN
UNIVERSITY OF KANSAS

A SAGE Reference Publication

 **SAGE Publications**
Thousand Oaks ■ London ■ New Delhi

Copyright © 2007 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

For information:



SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of measurement and statistics / editor Neil J. Salkind.
p. cm.

A SAGE Reference Publication.

Includes bibliographical references and index.

ISBN 1-4129-1611-9 or 978-1-4129-1611-0 (cloth)

1. Social sciences—Statistical methods—Encyclopedias. 2. Social sciences—Research—Methodology—Encyclopedias.

I. Salkind, Neil J.

HA29.S2363 2007

001.403—dc22

2006011888

This book is printed on acid-free paper.

06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

<i>Publisher:</i>	Rolf Janke
<i>Acquisitions Editor:</i>	Lisa Cuevas Shaw
<i>Reference Systems Coordinator:</i>	Leticia Gutierrez
<i>Project Editor:</i>	Tracy Alpern
<i>Copy Editors:</i>	Bonnie Freeman Liann Lech Carla Freeman
<i>Typesetter:</i>	C&M Digital (P) Ltd.
<i>Indexer:</i>	David Luljak
<i>Cover Designer:</i>	Michelle Kenny

Contents

Editorial Board, *vi*
List of Entries, *vii*
Reader's Guide, *xiii*
About the Editor, *xxi*
Contributors, *xxii*
Preface, *xxxiii*

Entries

Volume I: A–G
1–424

Volume II: H–P
425–798

Volume III: Q–Z
799–1058

Appendix A

Ten Commandments of Data Collection, *1059*

Tables of Critical Values, *1061*

Appendix B

Internet Sites About Statistics, *1073*

Appendix C

Glossary, *1081*

Master Bibliography, *1087–1136*

Index, *I-1–I-57*

Editorial Board

Editor

Neil J. Salkind
University of Kansas

Managing Editor

Kristin Rasmussen
University of Kansas

Advisory Board

Jeffrey Banfield
Department of Mathematical Sciences
Montana State University

Bruce Frey
Department of Psychology and Research in Education
University of Kansas

Wenxin Jiang
Department of Statistics
Northwestern University

Galin L. Jones
School of Statistics
University of Minnesota

Jianan Peng
Department of Mathematics and Statistics
Acadia University

Jerome P. Reiter
Practice of Statistics and Decision Sciences
Institute of Statistics and Decision Sciences
Duke University

List of Entries

Ability Tests	Audit Trail
Abstracts	Authenticity
Acceptance Sampling	Autocorrelation
Achievement Tests	Average
Active Life Expectancy	Average Deviation
Adaptive Sampling Design	
Adjective Checklist	Babbage, Charles
Age Norms	Bar Chart
Akaike Information Criterion	Basal Age
Alcohol Use Inventory	Basic Personality Inventory
Alternate Assessment	Basic Research
Alternative Hypothesis	Bayes Factors
American Doctoral Dissertations	Bayesian Information Criterion
American Psychological Association	Bayesian Statistics
American Psychological Society. <i>See</i> Association for Psychological Science	Bayley Scales of Infant Development
American Statistical Association	Beck Depression Inventory
Americans with Disabilities Act	Behavior Assessment System for Children
Analysis of Covariance (ANCOVA)	Behrens-Fisher Test
Analysis of Variance (ANOVA)	Bender Visual Motor Gestalt Test
Anthropometry	Bernoulli, Jakob
Applied Research	Binomial Distribution/Binomial and Sign Tests
Aptitude Tests	Binomial Test
Area Chart	Bioinformatics
Arithmetic Mean	Biserial Correlation Coefficients
Armed Forces Qualification Test	Bivariate Distributions
Armed Services Vocational Aptitude Battery	Bonferroni, Carlo Emilio
Artificial Neural Network	Bonferroni Test
Assessment of Interactions in Multiple Regression	Bowker Procedure
Association for Psychological Science	Box Plot (Box and Whisker Plot)
Asymmetry of g	Bracken Basic Concept Scale–Revised
Attenuation, Correction for	Bruno, James Edward
Attitude Tests	Buros Institute of Mental Measurements
Attributable Risk	
Attrition Bias	California Psychological Inventory
	Career Assessment Inventory

- Career Development Inventory
- Career Maturity Inventory
- Carroll Depression Scale
- Categorical Variable
- Causal Analysis
- Censored Data
- Centers for Disease Control and Prevention
- Central Limit Theorem
- Centroid
- Chance
- Chi-Square Test for Goodness of Fit
- Chi-Square Test for Independence
- Children's Academic Intrinsic Motivation Inventory
- Class Interval
- Classical Test Theory
- Classification and Regression Tree
- Clinical Assessment of Attention Deficit
- Clinical Assessment of Behavior
- Clinical Assessment of Depression
- Cluster Analysis
- Cluster Sampling
- Cochran Q Test
- Coefficient Alpha
- Coefficients of Correlation, Alienation, and Determination
- Cognitive Abilities Test
- Cognitive Psychometric Assessment
- Cohen's Kappa
- Complete Independence Hypothesis
- Completion Items
- Computational Statistics
- Computerized Adaptive Testing
- Comrey, Andrew L.
- Comrey Personality Scales
- Conditional Probability
- Confidence Intervals
- Construct Validity
- Content Validity
- Continuous Variable
- Contour Plot
- Convenience Sampling
- Coping Resources Inventory for Stress
- Correlation Coefficient
- Correspondence Analysis
- Covariance
- Criterion-Referenced Test
- Criterion Validity
- Critical Value
- Cronbach, Lee J.
- Culture Fair Intelligence Test
- Cumulative Frequency Distribution
- Curriculum-Based Measurement
- Curse of Dimensionality
- Curvilinear Regression
- Darwin, Charles
- Data Analysis ToolPak
- Data Collection
- Data Compression
- Data Mining
- Decision Boundary
- Decision Theory
- Delphi Technique
- Delta Method
- Deming, William Edwards
- Dependent Variable
- Descriptive Research
- Deviation Score
- Diagnostic Validity
- Difference Score
- Differential Aptitude Test
- Diggle-Kenward Model for Dropout
- Dimension Reduction
- Discriminant Analysis
- Discriminant Correspondence Analysis
- Dissimilarity Coefficient
- Distance
- DISTATIS
- Dixon Test for Outliers
- Dunn's Multiple Comparison Test
- Ecological Momentary Assessment
- Educational Testing Service
- Edwards Personal Preference Schedule
- Effect Size
- Eigendecomposition
- Eigenvalues
- EM Algorithm
- Embedded Figures Test
- Equivalence Testing
- Essay Items
- Estimates of the Population Median
- Ethical Issues in Testing

-
- Ethical Principles in the Conduct of Research With Human Participants
Evidence-Based Practice
Excel Spreadsheet Functions
Exploratory Data Analysis
Exploratory Factor Analysis
Eyeball Estimation
- Face Validity
Factor Analysis
Factor Scores
Factorial Design
Fagan Test of Infant Intelligence
Family Environment Scale
File Drawer Problem
Fisher, Ronald Aylmer
Fisher Exact Probability Test
Fisher-Irwin Test. *See* Fisher Exact Probability Test
Fisher's LSD
Fisher's Z Transformation
Fourier Transform
Fractal
Fractional Randomized Block Design
Frequency Distribution
Friedman Test
- Galton, Sir Francis
Gambler's Fallacy
Gauss, Carl Friedrich
Generalized Additive Model
Generalized Estimating Equations
Generalized Method of Moments
Generalized Procrustes Analysis
Gerontological Apperception Test
Gf-Gc Theory of Intelligence
Goodenough Harris Drawing Test
Goodness-of-Fit Tests
Graduate Record Examinations
Grand Mean
Graphical Statistical Methods
Gresham, Frank M.
Grounded Theory
Guttman Scaling
- Harmonic Mean
Health Insurance Portability and Accountability Act
Hello-Goodbye Effect
- Heteroscedasticity and Homoscedasticity
Hierarchical Linear Modeling
High-Stakes Tests
Histogram
Historiometrics
Holden Psychological Screening Inventory
Homogeneity of Variance
Hypergeometric Distribution
Hypothesis and Hypothesis Testing
- Illinois Test of Psycholinguistic Abilities
Immediate and Delayed Memory Tasks
Independent Variable
Individuals with Disabilities Education Act
Inferential Statistics
Information Referenced Testing
Information Systems Interaction Readiness Scales
Informed Consent
Instrumental Variables
Intelligence Quotient
Intelligence Tests
Internal External Locus of Control Scale
Internal Review Board
International Assessment of Educational Progress
Interrater Reliability
Interval Level of Measurement
Iowa Tests of Basic Skills
Iowa Tests of Educational Development
Ipsative Measure
Item and Test Bias
Item Response Theory
- Jackson, Douglas N.
Jackson Personality Inventory–Revised
Jackson Vocational Interest Survey
Journal of the American Statistical Association
Journal of Modern Applied Statistical Methods
Journal of Statistics Education
- k-Means Cluster Analysis
Kaufman Assessment Battery for Children
Kendall Rank Correlation
Kinetic Family Drawing Test
Kingston Standardized Cognitive Assessment
Kolmogorov-Smirnov Test for One Sample
Kolmogorov-Smirnov Test for Two Samples
KR-20 and KR-21

Kruskal-Wallis One-Way Analysis of Variance
Kuder Occupational Interest Survey
Kurtosis

Laboratory Behavioral Measures of Impulsivity
Latent Class Analysis
Law of Large Numbers
Law School Admissions Test
Least Squares, Method of
Life Values Inventory
Likelihood Ratio Test
Likert Scaling
Lilliefors Test for Normality
Line Chart
Linear Regression
Logistic Regression Analysis
Loglinear Analysis
Longitudinal/Repeated Measures Data
Luria Nebraska Neuropsychological Battery

Male Role Norms Inventory
Malthus, Thomas
Mann-Whitney U Test (Wilcoxon Rank-Sum Test)
Markov, Andrei Andreevich
Markov Chain Monte Carlo Methods
Matrix Analogies Test
Matrix Operations
Maximum Likelihood Method
McNemar Test for Significance of Changes
Mean
Measurement
Measurement Error
Measures of Central Tendency
Median
Median Test
Meta-Analysis
Metric Multidimensional Scaling
Millon Behavioral Medicine Diagnostic
Millon Clinical Multiaxial Inventory-III
Minnesota Clerical Test
Minnesota Multiphasic Personality Inventory
Missing Data Method
Mixed Models
Mixture Models
Mixtures of Experts
Mode

Moderator Variable
Monte Carlo Methods
Mosaic Plots
Moving Average
Multicollinearity
Multidimensional Aptitude Battery
Multiple Affect Adjective Checklist-Revised
Multiple-Choice Items
Multiple Comparisons
Multiple Correlation Coefficient
Multiple Correspondence Analysis
Multiple Factor Analysis
Multiple Imputation for Missing Data
Multitrait Multimethod Matrix and Construct Validity
Multivariate Analysis of Variance (MANOVA)
Multivariate Normal Distribution
Myers-Briggs Type Indicator

National Council on Measurement in Education
National Science Foundation
NEO Personality Inventory
Neonatal Behavioral Assessment Scale
Newman-Keuls Test
Nominal Level of Measurement
Nomothetic Versus Idiographic
Nonparametric Statistics
Nonprobability Sampling
Normal Curve
Null Hypothesis Significance Testing

O'Brien Test for Homogeneity of Variance
Observational Studies
Ockham's Razor
Ogive
One- and Two-Tailed Tests
One-Way Analysis of Variance
Ordinal Level of Measurement
Orthogonal Predictors in Regression

Page's *L* Test
Paired Samples *t* Test (Dependent Samples *t* Test)
Pairwise Comparisons
Parallel Coordinate Plots
Parallel Forms Reliability
Parameter

-
- Parameter Invariance
Part and Partial Correlations
Partial Least Square Regression
Pascal, Blaise
Path Analysis
Peabody Picture Vocabulary Test
Pearson, Karl
Pearson Product-Moment Correlation Coefficient
Percentile and Percentile Rank
Performance IQ
Performance-Based Assessment
Peritz Procedure
Personal Projects Analysis
Personality Assessment Inventory
Personality Research Form
Personality Tests
Pie Chart
Piers-Harris Children's Self-Concept Scale
Poisson, Siméon Denis
Poisson Distribution
Portfolio Assessment
Post Hoc Comparisons
Posterior Distribution
Predictive Validity
Preschool Language Assessment Instrument
Principal Component Analysis
Prior Distribution
Probability Sampling
Profile Analysis
Projective Hand Test
Projective Testing
Propensity Scores
Psychological Abstracts
Psychometrics
PsycINFO

Q Methodology
Q-Q Plot
Quality of Well-Being Scale
Quasi-Experimental Method
Questionnaires
Quota Sampling

Random Numbers
Random Sampling
Range

Rasch Measurement Model
Ratio Level of Measurement
Raven's Progressive Matrices
Record Linkage
Regression Analysis
Relative Risk
Reliability Theory
Repeated Measures Analysis of Variance
Residuals
Response to Intervention
Reverse Scaling
Reynolds, Cecil R.
Roberts Apperception Test for Children
Rorschach Inkblot Test
 R_V and Congruence Coefficients

Sample
Sample Size
Sampling Distribution of a Statistic
Sampling Error
Scaling
Scan Statistic
Scattergram
Scree Plot
Secondary Data Analysis
Section 504 of the Rehabilitation Act of 1973
Self-Report
Semantic Differential
Semantic Differential Scale
Semi-Interquartile Range
Shapiro-Wilk Test for Normality
Signal Detection Theory
Significance Level
Simple Main Effect
Simpson's Paradox
Simpson's Rule
Simulated Annealing
Simulation Experiments
Single-Subject Designs
Singular and Generalized Singular Value
Decomposition
Six Sigma
Sixteen Personality Factor Questionnaire
Skewness
Smoothing
Social Climate Scales

- Social Skills Rating System
- Society for Research in Child Development
- Sociological Abstracts
- Spatial Learning Ability Test
- Spatial Statistics
- Spearman's Rho
- Split Half Reliability
- Spreadsheet Functions
- Spurious Correlation
- Standard Deviation
- Standard Error of the Mean
- Standard Error of Measurement
- Standard Scores
- Standards for Educational and Psychological Testing*
- Stanford Achievement Test
- Stanford-Binet Intelligence Scales
- Stanine
- STATIS
- Statistical Significance
- Stem-and-Leaf Display
- Stratified Random Sampling
- Strong Interest Inventory
- Stroop Color and Word Test
- Structural Equation Modeling
- Structured Clinical Interview for *DSM-IV*
- Sunflower Plot
- Support Vector Machines
- Suppressor Variable
- Survey Weights
- Survival Analysis
- System of Multicultural Pluralistic Assessment
- T* Scores
- t* Test for Two Population Means
- Test-Retest Reliability
- Tests of Mediating Effects
- Text Analysis
- Thematic Apperception Test
- Three-Card Method
- Thurstone Scales
- Time Series Analysis
- Torrance, E. Paul
- Torrance Tests of Creative Thinking
- Torrance Thinking Creatively in Action and Movement
- Tree Diagram
- True/False Items
- True Score
- Tukey-Kramer Procedure
- Type I Error
- Type II Error
- Unbiased Estimator
- Universal Nonverbal Intelligence Test
- Validity Coefficient
- Validity Theory
- Variable
- Variable Deletion
- Variance
- Verbal IQ
- Vineland Adaptive Behavior Scales
- Vineland Social Maturity Scale
- Wechsler Adult Intelligence Scale
- Wechsler Individual Achievement Test
- Wechsler Preschool and Primary Scale of Intelligence
- West Haven-Yale Multidimensional Pain Inventory
- Wilcoxon, Frank
- Wilcoxon Mann-Whitney Test. *See* Mann-Whitney U Test (Wilcoxon Rank-Sum Test)
- Wilcoxon Signed Ranks Test
- Woodcock Johnson Psychoeducational Battery
- Woodcock Reading Mastery Tests Revised
- z* Scores

Reader's Guide

The purpose of the Reader's Guide is to provide you with a tool you can use to locate specific entries in the encyclopedia, as well as to find out what other related entries might be of interest to you. For example, if you are interested in the visual display of information and want to learn how to create a bar chart (under the general heading of Charts, Graphs, and Visual Displays in the Reader's Guide), you can also find reference to such entries as Histogram, Line Chart, and Mosaic Plots, all related to the same general topic.

The Reader's Guide is also a very direct and simple way to get an overview of which items are contained in the encyclopedia. Although each of the categories lists items in alphabetic order (as the encyclopedia is organized), you can glance through the main headings of the guide and then focus more on a particular area of interest. Then, just turn to any particular entry you want to locate. These are easily found because they appear in alphabetical order.

Biographies

Babbage, Charles
Bernoulli, Jakob
Bonferroni, Carlo Emilio
Bruno, James Edward
Comrey, Andrew L.
Cronbach, Lee J.
Darwin, Charles
Deming, William Edwards
Fisher, Ronald Aylmer
Galton, Sir Francis
Gauss, Carl Friedrich
Gresham, Frank M.
Jackson, Douglas N.
Malthus, Thomas
Markov, Andrei Andreevich
Pascal, Blaise
Pearson, Karl
Poisson, Siméon Denis
Reynolds, Cecil R.

Torrance, E. Paul
Wilcoxon, Frank

Charts, Graphs, and Visual Displays

Area Chart
Bar Chart
Box Plot (Box and Whisker Plot)
Contour Plot
Eyeball Estimation
Frequency Distribution
Histogram
Line Chart
Mosaic Plots
Ogive
Parallel Coordinate Plots
Pie Chart
Q-Q Plot
Scattergram
Scree Plot
Smoothing

Stem-and-Leaf Display
Sunflower Plot
Tree Diagram

Computer Topics and Tools

Babbage, Charles
Computational Statistics
Computerized Adaptive Testing
Curvilinear Regression
Data Analysis ToolPak
Data Compression
DISTATIS
Excel Spreadsheet Functions
Linear Regression
Residuals
Spatial Statistics
Spreadsheet Functions
STATIS

Concepts and Issues in Measurement

Ability Tests
Achievement Tests
Alternate Assessment
Americans with Disabilities Act
Anthropometry
Aptitude Tests
Artificial Neural Network
Asymmetry of g
Attitude Tests
Basal Age
Categorical Variable
Classical Test Theory
Coefficient Alpha
Completion Items
Computerized Adaptive Testing
Construct Validity
Content Validity
Criterion-Referenced Test
Criterion Validity
Cronbach, Lee J.
Curriculum-Based Measurement
Diagnostic Validity
Educational Testing Service
Equivalence Testing

Essay Items
Ethical Issues in Testing
Face Validity
Gf-Gc Theory of Intelligence
Guttman Scaling
Health Insurance Portability and Accountability Act
High-Stakes Tests
Immediate and Delayed Memory Tasks
Individuals with Disabilities Education Act
Information Referenced Testing
Informed Consent
Intelligence Quotient
Intelligence Tests
Internal Review Board
Interrater Reliability
Interval Level of Measurement
Ipsative Measure
Item and Test Bias
Item Response Theory
KR-20 and KR-21
Likert Scaling
Measurement
Measurement Error
Metric Multidimensional Scaling
Multiple-Choice Items
Multitrait Multimethod Matrix and Construct Validity
Nomothetic Versus Idiographic
Ordinal Level of Measurement
Parallel Forms Reliability
Performance IQ
Performance-Based Assessment
Personality Tests
Portfolio Assessment
Predictive Validity
Projective Testing
Q Methodology
Questionnaires
Ratio Level of Measurement
Reliability Theory
Response to Intervention
Reverse Scaling
Scaling
Section 504 of the Rehabilitation Act of 1973
Self-Report
Semantic Differential

Semantic Differential Scale
Six Sigma
Spearman's Rho
Split Half Reliability
Standard Error of Measurement
Standard Scores
Standards for Educational and Psychological Testing
T Scores
Test-Retest Reliability
Thurstone Scaling
Torrance, E. Paul
True/False Items
Validity Coefficient
Validity Theory
Verbal IQ
z Scores

Concepts and Issues in Statistics

Artificial Neural Network
Attenuation, Correction for
Autocorrelation
Bayesian Statistics
Bioinformatics
Central Limit Theorem
Decision Theory
Diggle-Kenward Model for Dropout
DISTATIS
Exploratory Factor Analysis
Factorial Design
Fourier Transform
Generalized Additive Model
Generalized Method of Moments
Generalized Procrustes Analysis
Graphical Statistical Methods
Hierarchical Linear Modeling
Historiometrics
Logistic Regression Analysis
Loglinear Analysis
Markov Chain Monte Carlo Methods
Matrix Operations
Mean
Measurement Error
Mixtures of Experts
Nonparametric Statistics
Propensity Scores

Rasch Measurement Model
Regression Analysis
Sampling Distribution of a Statistic
Signal Detection Theory
Simpson's Paradox
Spurious Correlation
Standard Error of the Mean
Standard Scores
Support Vector Machines
Survival Analysis
Type I Error
Type II Error

Data and Data Reduction Techniques

Censored Data
Data Compression
Data Mining
Discriminant Analysis
Eigenvalues
Exploratory Data Analysis
Factor Analysis
Factor Scores
Missing Data Method
Multiple Factor Analysis
Record Linkage
Secondary Analysis of Data

Descriptive Statistics

Arithmetic Mean
Attenuation, Correction for
Autocorrelation
Average
Average Deviation
Bayley Scales of Infant Development
Biserial Correlation Coefficient
Class Interval
Coefficients of Correlation, Alienation, and
Determination
Cognitive Psychometric Assessment
Cohen's Kappa
Correlation Coefficient
Cumulative Frequency Distribution
Deviation Score
Difference Score

Estimates of the Population Median
Fisher's Z Transformation
Frequency Distribution
Galton, Sir Francis
Grand Mean
Harmonic Mean
Histogram
Kendall Rank Correlation
Mean
Measures of Central Tendency
Median
Mode
Moving Average
Parameter
Parameter Invariance
Part and Partial Correlations
Pearson, Karl
Pearson Product-Moment Correlation Coefficient
Percentile and Percentile Rank
 R_V and Congruence Coefficients
Scattergram
Semi-Interquartile Range
Spurious Correlation
Standard Deviation
Survey Weights
Text Analysis

Evaluation

Achievement Tests
Evidence-Based Practices
Health Insurance Portability and Accountability Act
High-Stakes Tests
Questionnaires

Experimental Methods

Alternative Hypothesis
American Statistical Association
Americans with Disabilities Act
Association for Psychological Science
Basic Research
Bioinformatics
Complete Independence Hypothesis
Continuous Variable
Critical Value
Data Collection

Data Mining
Delphi Technique
Dependent Variable
Descriptive Research
Ethical Issues in Testing
Ethical Principles in the Conduct of Research With Human Participants
Fractional Randomized Block Design
Hello-Goodbye Effect
Hypothesis and Hypothesis Testing
Independent Variable
Informed Consent
Instrumental Variables
Internal Review Board
Longitudinal/Repeated Measures Data
Meta-Analysis
Missing Data Method
Mixed Models
Mixture Models
Moderator Variable
Monte Carlo Methods
Null Hypothesis Significance Testing
Ockham's Razor
Pairwise Comparisons
Post Hoc Comparisons
Projective Testing
Quasi-Experimental Method
Sample Size
Section 504 of the Rehabilitation Act of 1973
Significance Level
Simple Main Effect
Simulation Experiments
Single-Subject Designs
Standards for Educational and Psychological Testing
Statistical Significance
Suppressor Variable
Variable
Variable Deletion
Variance

Inferential Statistics

Akaike Information Criterion
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Bayes Factors
Bayesian Information Criterion

Binomial Test
Bonferroni, Carlo Emilio
Complete Independence Hypothesis
Data Analysis ToolPak
Exploratory Factor Analysis
Factorial Design
Fisher, Ronald Aylmer
Hierarchical Linear Modeling
Hypothesis and Hypothesis Testing
Inferential Statistics
Logistic Regression Analysis
Markov, Andrei Andreevich
Null Hypothesis Significance Testing
Pairwise Comparisons
Part and Partial Correlations
Repeated Measures Analysis of Variance
Type I Error
Type II Error
Wilcoxon, Frank

Organizations and Publications

Abstracts
American Doctoral Dissertations
American Psychological Association
American Statistical Association
Association for Psychological Science
Buros Institute of Mental Measurements
Centers for Disease Control and Prevention
Educational Testing Service
Journal of the American Statistical Association
Journal of Modern Applied Statistical Methods
Journal of Statistics Education
National Science Foundation
Psychological Abstracts
Psychometrics
PsycINFO
Society for Research in Child Development
Sociological Abstracts

Prediction and Estimation

Attributable Risk
Bernoulli, Jakob
Chance
Conditional Probability
Confidence Intervals

Continuous Variable
Curse of Dimensionality
Decision Boundary
Decision Theory
File Drawer Problem
Gambler's Fallacy
Generalized Estimating Equations
Law of Large Numbers
Maximum Likelihood Method
Nonprobability Sampling
Pascal, Blaise
Probability Sampling
Random Numbers
Relative Risk
Signal Detection Theory
Significance Level
Three-Card Method

Probability

Alternate Assessment
Audit Trail
Authenticity
Categorical Variable
Essay Items
Grounded Theory
Observational Studies
Portfolio Assessment
Self-Report
Text Analysis

Qualitative Methods

Active Life Expectancy
Assessment of Interactions in Multiple Regression
Eyeball Estimation
Orthogonal Predictors in Regression
Regression Analysis
Survival Analysis

Samples, Sampling, and Distributions

Acceptance Sampling
Adaptive Sampling Design
Age Norms
Attrition Bias
Career Maturity Inventory

Central Limit Theorem
Class Interval
Cluster Sampling
Confidence Intervals
Convenience Sampling
Cumulative Frequency Distribution
Data Collection
Diggle-Kenward Model for Dropout
Gauss, Carl Friedrich
Heteroscedasticity and Homoscedasticity
Homogeneity of Variance
Hypergeometric Distribution
Kurtosis
Malthus, Thomas
Multicollinearity
Multivariate Normal Distribution
Nonprobability Sampling
Normal Curve
Ogive
Parameter
Percentile and Percentile Rank
Poisson, Siméon Denis
Poisson Distribution
Posterior Distribution
Prior Distribution
Probability Sampling
Quota Sampling
Random Sampling
Sample
Sample Size
Semi-Interquartile Range
Simpson's Rule
Skewness
Smoothing
Stanine
Stratified Random Sampling
Unbiased Estimator

Statistical Techniques

Binomial Distribution/Binomial and Sign Tests
Bivariate Distributions
Bonferroni Test
Bowker Procedure
Causal Analysis
Centroid

Chance
Chi-Square Test for Goodness of Fit
Chi-Square Test for Independence
Classification and Regression Tree
Cochran Q Test
Cohen's Kappa
Delta Method
Dimension Reduction
Discriminant Analysis
Dissimilarity Coefficient
Dixon Test for Outliers
Dunn's Multiple Comparison Test
Eigendecomposition
Eigenvalues
EM Algorithm
Exploratory Data Analysis
Factor Analysis
Factor Scores
Fisher Exact Probability Test
Fisher's LSD
Friedman Test
Goodness-of-Fit Tests
Grounded Theory
k-Means Cluster Analysis
Kolmogorov-Smirnov Test for One Sample
Kolmogorov-Smirnov Test for Two Samples
Kruskal-Wallis One-Way Analysis of Variance
Latent Class Analysis
Likelihood Ratio Test
Lilliefors Test for Normality
Mann-Whitney U Test (Wilcoxon Rank-Sum Test)
McNemar Test for Significance of Changes
Median Test
Meta-Analysis
Multiple Comparisons
Multiple Factor Analysis
Multiple Imputation for Missing Data
Multivariate Analysis of Variance (MANOVA)
Newman-Keuls Test
O'Brien Test for Homogeneity of Variance
Observational Studies
One-Way Analysis of Variance
Page's *L* Test
Paired Samples *t* Test (Dependent Samples *t* Test)
Path Analysis
Peritz Procedure

Scan Statistic
 Shapiro-Wilk Test for Normality
 Structural Equation Modeling
 t Test for Two Population Means
 Tests of Mediating Effects
 Three-Card Method
 Tukey-Kramer Procedure
 Wilcoxon Signed Ranks Test

Statistical Tests

Analysis of Covariance (ANCOVA)
 Analysis of Variance (ANOVA)
 Behrens-Fisher Test
 Binomial Distribution/Binomial and Sign Tests
 Binomial Test
 Bonferroni Test
 Bowker Procedure
 Chi-Square Test for Goodness of Fit
 Chi-Square Test for Independence
 Classification and Regression Tree
 Cochran Q Test
 Dixon Test for Outliers
 Dunn's Multiple Comparison Test
 Excel Spreadsheet Functions
 Fisher Exact Probability Test
 Fisher's LSD
 Friedman Test
 Goodness-of-Fit Tests
 Kolmogorov-Smirnov Test for One Sample
 Kolmogorov-Smirnov Test for Two Samples
 Kruskal-Wallis One-Way Analysis of Variance
 Latent Class Analysis
 Likelihood Ratio Test
 Lilliefors Test for Normality
 Mann-Whitney U Test (Wilcoxon Rank-Sum Test)
 McNemar Test for Significance of Changes
 Median Test
 Multiple Comparisons
 Multivariate Analysis of Variance (MANOVA)
 Newman-Keuls Test
 O'Brien Test for Homogeneity of Variance
 One- and Two-Tailed Tests
 One-Way Analysis of Variance
 Page's L Test
 Paired Samples t Test (Dependent Samples t Test)

Peritz Procedure
 Repeated Measures Analysis of Variance
 Shapiro-Wilk Test for Normality
 t Test for Two Population Means
 Tests of Mediating Effects
 Tukey-Kramer Procedure
 Wilcoxon Signed Ranks Test

Tests by Name

Adjective Checklist
 Alcohol Use Inventory
 Armed Forces Qualification Test
 Armed Services Vocational Aptitude Battery
 Basic Personality Inventory
 Bayley Scales of Infant Development
 Beck Depression Inventory
 Behavior Assessment System for Children
 Bender Visual Motor Gestalt Test
 Bracken Basic Concept Scale—Revised
 California Psychological Inventory
 Career Assessment Inventory
 Career Development Inventory
 Career Maturity Inventory
 Carroll Depression Scale
 Children's Academic Intrinsic Motivation Inventory
 Clinical Assessment of Attention Deficit
 Clinical Assessment of Behavior
 Clinical Assessment of Depression
 Cognitive Abilities Test
 Cognitive Psychometric Assessment
 Comrey Personality Scales
 Coping Resources Inventory for Stress
 Culture Fair Intelligence Test
 Differential Aptitude Test
 Ecological Momentary Assessment
 Edwards Personal Preference Schedule
 Embedded Figures Test
 Fagan Test of Infant Intelligence
 Family Environment Scale
 Gerontological Apperception Test
 Goodenough Harris Drawing Test
 Graduate Record Examinations
 Holden Psychological Screening Inventory
 Illinois Test of Psycholinguistic Abilities
 Information Systems Interaction Readiness

Internal External Locus of Control Scale	Profile Analysis
International Assessment of Educational Progress	Projective Hand Test
Iowa Tests of Basic Skills	Quality of Well-Being Scale
Iowa Tests of Educational Development	Raven's Progressive Matrices
Jackson Personality Inventory–Revised	Roberts Apperception Test for Children
Jackson Vocational Interest Survey	Rorschach Inkblot Test
Kaufman Assessment Battery for Children	Sixteen Personality Factor Questionnaire
Kinetic Family Drawing Test	Social Climate Scales
Kingston Standardized Cognitive Assessment	Social Skills Rating System
Kuder Occupational Interest Survey	Spatial Learning Ability Test
Laboratory Behavioral Measures of Impulsivity	Stanford Achievement Test
Law School Admissions Test	Stanford-Binet Intelligence Scales
Life Values Inventory	Strong Interest Inventory
Luria Nebraska Neuropsychological Battery	Stroop Color and Word Test
Male Role Norms Inventory	Structured Clinical Interview for <i>DSM-IV</i>
Matrix Analogies Test	System of Multicultural Pluralistic Assessment
Millon Behavioral Medicine Diagnostic	Thematic Apperception Test
Millon Clinical Multiaxial Inventory-III	Torrance Tests of Creative Thinking
Minnesota Clerical Test	Torrance Thinking Creatively in Action and Movement
Minnesota Multiphasic Personality Inventory	Universal Nonverbal Intelligence Test
Multidimensional Aptitude Battery	Vineland Adaptive Behavior Scales
Multiple Affect Adjective Checklist–Revised	Vineland Social Maturity Scale
Myers-Briggs Type Indicator	Wechsler Adult Intelligence Scale
NEO Personality Inventory	Wechsler Individual Achievement Test
Neonatal Behavioral Assessment Scale	Wechsler Preschool and Primary Scale of Intelligence
Peabody Picture Vocabulary Test	West Haven-Yale Multidimensional Pain Inventory
Personal Projects Analysis	Woodcock Johnson Psychoeducational Battery
Personality Assessment Inventory	Woodcock Reading Mastery Tests Revised
Personality Research Form	
Piers-Harris Children's Self-Concept Scale	
Preschool Language Assessment Instrument	

About the Editor

Neil J. Salkind (PhD, University of Maryland, 1973) is a Professor of Psychology and Research in Education at the University of Kansas in Lawrence, Kansas. He completed postdoctoral training as a member of the Bush Child and Family Policy Institute at the University of North Carolina and has authored and coauthored more than 125 scholarly papers and books. Most recently, he is the author of *Statistics for People Who (Think They) Hate Statistics: The Excel Edition* (2007), *Tests & Measurement for People Who*

(Think They) Hate Tests & Measurement (2006), the *Encyclopedia of Human Development* (2006), *Theories of Human Development* (2004), and *Statistics for People Who (Think They) Hate Statistics* (2004), all published by Sage. He was the editor of *Child Development Abstracts and Bibliography*, published by the Society for Research in Child Development (SRCD), from 1988 through 2001, and he is the treasurer-elect of Division 7 of the American Psychological Association.

Contributors

Francisco J. Abad
Universidad Autonoma de Madrid

Inmaculada Aban
University of Alabama at Birmingham

Hervé Abdi
University of Texas at Dallas

Phillip L. Ackerman
Georgia Institute of Technology

Demetrios S. Alexopoulos
University of Patras, Greece

Audrey Amrein-Beardsley
Arizona State University

Lauren E. Auld
DePauw University

Carrie R. Ball
University of Wisconsin–Madison

Kimberly A. Barchard
University of Nevada, Las Vegas

Jonathan Barzilai
Dalhousie University

Edward J. Bedrick
University of New Mexico

Mark L. Berenson
Montclair State University

Dongsheng Bi
University of Nebraska Lincoln

Damian P. Birney
University of Sydney

David M. Boynton
Saint Michael's College

Bruce A. Bracken
College of William & Mary

Jennifer Bragger
Montclair State University

Gary G. Brannigan
State University of New York–Plattsburgh

Ernest W. Brewer
University of Tennessee

Carolyn Brodbeck
Chapman University

Sarah Brookhart
American Psychological Society

Duane Brown
University of North Carolina, Chapel Hill

Jennifer Ann Brown
University of Canterbury

Shawn T. Bubany
University of Minnesota

Michael J. Burke
Tulane University

Mary Margaret Capraro
Texas A&M University

Robert M. Capraro
Texas A&M University

Joseph E. Cavanaugh
University of Iowa

Hua-Hua Chang
University of Illinois

Elaine Chapman
University of Western Australia

Bernard C. K. Choi
Public Health Agency of Canada

Siu L. Chow
University of Regina

Michelle D. Chudleigh
Alberta Hospital Edmonton

Moo K. Chung
University of Wisconsin

M. H. Clark
Southern Illinois University

Murray Clayton
University of Wisconsin–Madison

A. Jill Clemence
Austen Riggs Center

Roberto Colom
Universidad Autonoma de Madrid

John Colombo
University of Kansas

Andrew L. Comrey
University of California, Los Angeles

Dianne Cook
Iowa State University

R. Kelly Crace
College of William & Mary

Bonnie Cramond
University of Georgia

William L. Curlette
Georgia State University

Larry Daniel
University of North Florida

Craig Darch
University of Auburn

Duncan Day
Queen's University

E. Jacquelin Dietz
Meredith College

Bryan J. Dik
Colorado State University

Heather Doescher
University of Wisconsin–Madison

Romilia Domínguez de Ramírez
University of Houston

Joseph A. Doster
University of North Texas

Donald M. Dougherty
Wake Forest University Baptist Medical Center

Ronald C. Eaves
Auburn University

Anna Ebel-Lam
Queen's University

Maeghan N. Edwards
Pennsylvania State University

Stephen N. Elliott
Vanderbilt University

Susan Embretson
Georgia Institute of Technology

Craig K. Enders
Arizona State University

Marie Joelle Estrada
University of North Carolina, Chapel Hill

Martin G. Evans
University of Toronto

Leandre R. Fabrigar
Queen's University

Gail F. Fahoome
Wayne State University

Andy P. Field
University of Sussex

Barry Forer
University of British Columbia

Robert A. Forsyth
University of Iowa

Brian F. French
Purdue University

Kathryn H. Ganske
Georgia State University

Travis L. Gee
University of Queensland, Australia

Carole E. Gelfer
William Paterson University

James E. Gentle
George Mason University

Morton A. Gernsbacher
University of Wisconsin–Madison

Maribeth Gettinger
University of Wisconsin–Madison

Marjan Ghahramanlou-Holloway
Uniformed Services University of the Health Sciences

Lisa M. Given
University of Alberta

Kevin W. Glavin
Kent State University

Charles Golden
Nova Southeastern University

Adele Eskeles Gottfried
California State University, Northridge

Naomi Grant
Queen's University

James W. Grice
Oklahoma State University

Erik J. Groessl
VA San Diego / University of California, San Diego

Lyndsi M. Grover
University of North Texas

Suzanne M. Grundy
California State University, San Bernardino

Anthony J. Guarino
University of Auburn

Mads Haahr
Trinity College Dublin

John W. Hagen
Society of Research in Child Development

Brian Haig
University of Canterbury

Thomas Haladyna
Arizona State University

Young-Hoon Ham
University of Tennessee

Ronald K. Hambleton
University of Massachusetts

Chirok Han
Victoria University of Wellington

David Hann
University of Kansas

Jo-Ida C. Hansen
University of Minnesota

James W. Hardin
University of South Carolina

Clay Helberg
SPSS Inc.

Susanne Hempel
University of York

Ryan G. Henry
Brigham Young University

Kristin Heron
Syracuse University

Matthew J. Hertenstein
DePauw University

Christine L. Himes
Syracuse University

Nathaniel R. Hirtz
Murray State University

David B. Hitchcock
University of South Carolina

James S. Ho
Alberta Hospital Edmonton

Heike Hofmann
Iowa State University

Cody S. Hollist
University of Nebraska–Lincoln

Johnny Holloway
American University

Robert Hopkins
Queens University

Jennifer R. Hsu
William Paterson University

Louis M. Hsu
Fairleigh Dickinson University

Allison Huck
University of Kentucky

Schuyler W. Huck
University of Tennessee

Bradley E. Huitema
Western Michigan University

Russell T. Hurlburt
University of Nevada, Las Vegas

Jennifer Ivie
University of Kansas

Robert A. Jacobs
University of Rochester

John Jamieson
Lakehead University

Galin L. Jones
University of Minnesota

Samuel Juni
New York University

Sema A. Kalaian
Eastern Michigan University

Matthew E. Kaler
University of Minnesota

Kristen M. Kalymon
University of Wisconsin–Madison

Robert M. Kaplan
University of California, Los Angeles

Michael A. Karchmer
Gallaudet Research Institute

Michael Karson
University of Denver

Rafa M. Kasim
Kent State University

Allison B. Kaufman
University of California, Riverside

James C. Kaufman
California State University, San Bernardino

Lisa Keller
University of Massachusetts, Amherst

Lindy Kilik
Queen's University

Kyung Hee Kim
Eastern Michigan University

Roger E. Kirk
Baylor University

Steve Kirkland
University of Regina

Theresa Kline
University of Calgary

James Randolph Knaub, Jr.
U.S. Government, Energy Information Administration

John C. Kolar
Medical City Children's Hospital, Dallas

Nicole B. Koppel
Montclair State University

Richard M. Kreminski
Texas A&M University–Commerce

Joachim I. Krueger
Brown University

Thomas Kubiszyn
University of Houston

Jouni Kuha
London School of Economics

Jonna M. Kulikowich
Pennsylvania State University

Wilda Laija-Rodriguez
California State University, Northridge

David M. Lane
Rice University

Sean Laraway
San Jose State University

Michael D. Larsen
Iowa State University

Nicole Lazar
University of Georgia

Howard B. Lee
University of California, Riverside

W. Vanessa Lee
University of Minnesota

Nancy L. Leech
Colorado University, Denver

Lawrence Leemis
College of William & Mary

Pui-Wa Lei
Pennsylvania State University

Russell V. Lenth
University of Iowa

Ming-Ying Leung
University of Texas at El Paso

Melanie Leuty
University of Minnesota

Ronald F. Levant
University of Akron

Robert W. Levenson
University of California, Berkeley

Jacob J. Levy
University of Tennessee

Bruce Lindsay
Pennsylvania State University

Brian R. Little
Carleton University and Harvard University

David F. Lohman
University of Iowa

Jeffrey D. Long
University of Minnesota

Sarah A. Lopienski
Kent State University

Kathryn Lou
University of Pennsylvania

Gerard H. Maassen
Utrecht University

Karen MacGregor
Queens University

Effie Maclellan
University of Strathclyde

W. Todd Maddox
University of Texas at Austin

Silvia A. Madrid
University of Arizona

Susan J. Maller
Purdue University

Dawn M. Marsh
Wake Forest University Baptist Medical Center

Luci A. Martin
University of North Texas

Kenneth B. Matheny
Georgia State University

Charles W. Mathias
Wake Forest University Health Sciences

Sonia Matwin
University of Utah

Mary Ann McCabe
American Psychological Association

Geoffrey McLachlan
University of Queensland

Adam W. Meade
North Carolina State University

Christopher J. Mecklin
Murray State University

Franklin Mendivil
Acadia University

Jorge L. Mendoza
University of Oklahoma

George Michailidis
University of Michigan

Jeremy Miles
University of York

Richard B. Miller
Brigham Young University

Ross E. Mitchell
Gallaudet University

Amitava Mitra
Auburn University

Geert Molenberghs
University Hasselt

Paul Molin
University of Bourgogne

Dirk F. Moore
University of Medicine and Dentistry of New Jersey

Kevin E. Moore
DePaul University

Bernice S. Moos
Stanford University

Rudolf H. Moos
Stanford University

Mark Mostert
Regent University

Ronald Neath
University of Minnesota

Liqiang Ni
University of Central Florida

Adelheid A. M. Nicol
Royal Military College

Meghan E. Norris
Queen's University

Anthony J. Onwuegbuzie
University of South Florida

J. Shelly Paik
Queen's University

Anita W. P. Pak
University of Ottawa

Paul E. Panek
Ohio State University–Newark

Hans Anand Pant
Humboldt University of Berlin

Dong-Ho Park
University of Tennessee

Scott Parker
American University

Sandrine Pavoine
Muséum National d'Histoire Naturelle, Paris

Manohar Pawar
Charles Sturt University

Edsel Pena
University of South Carolina

Sarah Peterson
University of Kansas

Andrew M. Pomerantz
Southern Illinois University Edwardsville

Jennifer L. Porter
DePauw University

Ronald D. Porter
Queen's University

Patricia Ramsey
Fordham University

Philip H. Ramsey
Queens College of City University of New York

John Randal
Victoria University of Wellington

Kristin Rasmussen
University of Kansas

Marco Reale
University of Canterbury

John R. Reddon
Alberta Hospital Edmonton

Malcolm James Ree
Our Lady of the Lake University

Jerome Reiter
Duke University

Bixiang Ren
University of Tennessee

Alberto Restori
California State University, Northridge

James C. Rhoads
Westminster College

Andrew T. Roach
Vanderbilt University

Beth Rodgers
University of Wisconsin–Milwaukee

Michael C. Rodriguez
University of Minnesota

Ward Rodriguez
California State University, East Bay

Javier Rojo
Rice University

Patrick J. Rosopa
University of Central Florida

Thomas E. Rudy
University of Pittsburgh

André A. Rupp
Humboldt University

Charles J. Russo
University of Dayton

Thomas Rutledge
UC San Diego

Steve Saladin
University of Idaho

Neil J. Salkind
University of Kansas

Elizabeth M. Salter
University of Texas at Dallas

Mark L. Savickas
Northeast Ohio Universities College of Medicine

Shlomo S. Sawilowsky
Wayne State University

Khalid Sayood
University of Nebraska–Lincoln

Carl J. Scarrott
University of Canterbury

Stanley L. Sclove
University of Illinois at Chicago

Kyoungah See
Miami University

William R. Shadish
University of California, Merced

Ramalingam Shanmugam
Texas State University

Boris Shulkin
Wayne State University

Dean Keith Simonton
University of California, Davis

Gary N. Siperstein
University of Massachusetts, Boston

Stephen G. Sireci
University of Massachusetts, Amherst

William P. Skorupski
University of Kansas

Joshua Smyth
Syracuse University

Robert A. Spies
University of Nebraska–Lincoln

Christopher J. Sroka
Ohio State University

Douglas Steinley
University of Missouri–Columbia

Steven E. Stemler
Wesleyan University

Michael Stewart
University of Sydney

David W. Stockburger
United States Air Force Academy

Eugene F. Stone-Romero
University of Texas, San Antonio

Bryce F. Sullivan
Southern Illinois University Edwardsville

Jun Sun
Texas A&M University

Martin A. Tanner
Northwestern University

Christopher P. Terry
Syracuse University

Robert M. Thorndike
Western Washington University

Davood Tofghi
Arizona State University

Larry Toothaker
University of Oklahoma

Marietta J. Tretter
Texas A&M University

George C. Tseng
University of Pittsburgh

Ping-Lun Tu
University of Tennessee

Jung-Ying Tzeng
North Carolina State University

Graham Upton
University of Essex

Dominique Valentin
University of Bourgogne

Nicholas G. Velissaris
Society for Research in Child Development

Geert Verbeke
Katholieke Universiteit Leuven

Fran Vertue
Child and Family Psychology Centre

Abdus S. Wahed
University of Pittsburgh

Harald Walach
University College Northampton

Russell F. Waugh
Edith Cowan University

Ann M. Weber
University of Wisconsin–Madison

Gail Weems
University of Arkansas Little Rock

Kimberly Weems
North Carolina State University

Kirsten Wells
Kansas University

Shane M. Whipple
Alberta Hospital Edmonton

Rand R. Wilcox
University of Southern California

Todd J. Wilkinson
University of Minnesota

Siân E. Williams
Canterbury Christ Church University

Thomas O. Williams, Jr.
Virginia Polytechnic Institute

Jay K. Wood
Queens University

Suzanne Woods-Groves
Auburn University

Daniel B. Wright
University of Sussex

Karl L. Wuensch
East Carolina University

Hongwei Yang
University of Tennessee–Knoxville

Keming Yang
University of Reading

Zhiliang Ying
Columbia University

Vincent R. Zalcik
Alberta Hospital Edmonton

April L. Zenisky
University of Massachusetts, Amherst

Jin Zhang
University of Manitoba

Zhigang Zhang
Oklahoma State University

Shuangmei (Christine) Zhou
University of Minnesota

Marvin Zuckerman
University of Delaware

Bruno D. Zumbo
University of British Columbia

Preface

It's an interesting paradox when an important subject, which can help us make sense of our busy, everyday world, is considered very difficult to approach. Such is the case with measurement and statistics. However, this does not necessarily have to be the case, and we believe that the *Encyclopedia of Measurement and Statistics* will show you why.

These two areas of study encompass a very wide range of topics, and a knowledge of even the basic concepts and ideas allows us to be much better prepared as intelligent consumers of information.

Whether we are interested in knowing if there is a difference between two groups in their preference for a particular brand of cereal or how the Americans with Disabilities Act works, we need to know how to analyze and interpret information. And often, when that information is in the form of numbers, that's where statistics and measurement come into play. That basic stat course in college might have been a nightmare, but the material is no more difficult to grasp and apply than is any other discipline in the social and behavioral sciences.

Although hundreds of books have been written about the different topics that are contained in the *Encyclopedia of Measurement and Statistics*, and there are thousands upon thousands of studies that have been conducted in this area, what we offer here is something quite different—a comprehensive overview of important topics. What we hope we have accomplished are entries that comprise a comprehensive overview of the most important topics in the areas of measurement and statistics—entries that share this important information in a way that is informative; not too technical; and even, in some cases, entertaining.

Through almost 500 contributions and some special elements that will be described later in this preface, experts in each of the entries contained in these pages contribute an overview and an explanation of the major topics in these two fields.

The underlying rationale for the selection of particular topics and their presentation in this encyclopedia comes from the need to share with the educated reader topics that are rich, diverse, and deserving of closer inspection. Within these pages, we provide the overview and the detail that we feel is necessary to become well acquainted with these topics.

As in many other encyclopedias, the *Encyclopedia of Measurement and Statistics* is organized in alphabetical order, from A through Z. However, particular themes were identified early on that could be used to organize conceptually the information and the entries. These themes or major topic areas constitute the Reader's Guide, which appears on page xiii. Categories such as Experimental Methods, Qualitative Methods, and Organizations and Publications are only a few of the many that help organize the entire set of contributions.

The Process

The first task in the creation of a multivolume encyclopedia such as this is the development of a complete and thorough listing of the important topics in the disciplines of measurement and statistics. This process started with the identification of entries that the editor and advisory board thought were important to include. We tried to make sure that these entries included topics that would be of interest to a general readership,

but we wanted to exclude terms and ideas that were too highly technical or too far removed from the interests of the everyday reader. This list was reviewed several times until we felt that it was a comprehensive set of topics that best fit the vision for the encyclopedia.

Like many other disciplines, there is a great deal of overlap between different important concepts and ideas in measurement and statistics. For example, although there is an entry titled Descriptive Statistics (which is a general overview), there is much greater detail in the entries titled Mean and Median. That overlap is fine because it provides two different, and compatible, views of the same topic and can only help reinforce one's knowledge. We hope that the cross-references we provide will help the user understand this and get the most out of learning about any one idea, term, or procedure.

As expected, this list was edited and revised as we worked and as authors were recruited to write particular entries. Enthusiastic authors suggested new topics that might have been overlooked as well as removing topics that might have no appeal. All of these changes were taken into consideration as the final list was assembled.

The next step was to assign a length to a particular entry, which ranged from 500 words for simple definitions or biographies (such as the one for the Arithmetic Mean or Charles Babbage, respectively) to almost 4,000 words for longer, more in-depth exploration for topics (such as the entry on Aptitude Tests). In between were articles that were 1,500 and 2,000 words in length. (At times, authors asked that the length be extended because they had so much information they wanted to share and they felt that the limitation on space was unwarranted. In most cases, it was not a problem to allow such an extension.)

The final step was the identification of authors. This took place through a variety of mechanisms, including the identification of individuals based on the advisory board recommendations and/or the editor's professional and personal experiences, authors of journal articles and books who focused on a particular area directly related to the entry, and referrals from other individuals who were well known in the field.

Once authors were identified and invited, and once they confirmed that they could participate, they were sent detailed instructions and given a deadline for the submission of their entry. The results, as you well know by now, after editing, layout, and other production steps, are in your hands.

How to Use This Reference

We know the study of measurement and statistics can be less than inviting. But, as we mentioned at the beginning of this preface, and want to emphasize once again here, the ideas and tools contained in these pages are approachable and can be invaluable for understanding our very technical world and an increasing flow of information.

Although many of the ideas you read about in these pages are relatively recent, some are centuries old. Yet both kinds hold promise for your being able to better navigate the increasingly complex world of information we each face every day.

So, although many of us believe that this encyclopedia should only be consulted when a term or idea needs some clarification, why not take some time and just browse through the material and see what types of entries are offered and how useful you might find them?

As we wrote earlier, a primary goal of creating this set of volumes was to open up the broad discipline of measurement and statistics to a wider and more general audience than usual.

Take these books and find a comfortable seat in the library, browse through the topics, and read the ones that catch your eye. We're confident that you'll continue reading and looking for additional related entries, such as "Applying Ideas on Statistics and Measurement," where you can find examples of how these ideas are applied and, in doing so, learn more about whatever interests you.

Should you want to find a topic within a particular area, consult the Reader's Guide, which organizes entries within this two-volume set into one general category. Using this tool, you can quickly move to an area or a specific topic that you find valuable and of interest.

Finally, there other elements that should be of interest.

Appendix A is a guide to basic statistics for those readers who might like a more instructional, step-by-step presentation of basic concepts in statistics and measurement. It also includes a table of critical values used in hypothesis testing and an important part of any reference in this area. These materials are taken from *Statistics for People Who (Think They) Hate Statistics*, written by the editor and also published by Sage.

Appendix B represents a collection of some important and useful sites on the Internet that have additional information about measurement and statistics. Although such sites tend to remain stable, there may be some changes in the Internet address. If you cannot find the Web page using the address that is provided, then search for the name of the Web site using Google or another search engine.

Finally, Appendix C is a glossary of terms and concepts you will frequently come across in these volumes.

Acknowledgments

This has been a challenging and rewarding project. It was ambitious in scope because it tried to corral a wide and diverse set of topics within measurement and statistics into a coherent set of volumes.

First, thanks to the Advisory Board, a group of scholars in many different areas that took the time to review the list of entries and make invaluable suggestions as to what the reader might find valuable and how that topic should be approached. The Advisory Board members are very busy people who took the time to help the editor develop a list that is broad in scope and represents the most important topics in human development. You can see a complete list of who these fine people are on page vi.

My editor and my publisher at Sage Publications, Lisa Cuevas Shaw and Rolf Janke, respectively,

deserve a great deal of thanks for bringing this project to me and providing the chance to make it work. They are terrific people who provide support and ideas and are always there to listen. And perhaps best of all, they get things done.

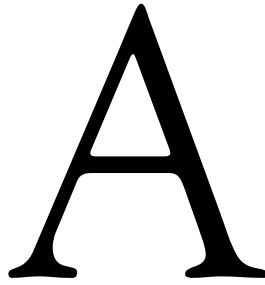
Other people also helped make this task enjoyable and helped create the useful, informative, and approachable set of volumes you hold in your hands. Among these people are Tracy Alpern, Sage senior project editor, and Bonnie Freeman, Liann Lech, and Carla Freeman, copy editors.

I'll save one of the biggest thank-yous for Kristin Rasmussen, the managing editor, who managed this project in every sense of the word, including the formidable tasks of tracking entries, submissions, reviews, and resubmissions. All of this was easily accomplished with enthusiasm, initiative, and perseverance when answering endless questions through thousands of e-mails to hundreds of authors. She is currently a doctoral student at the University of Kansas and has an exceptionally bright future. Thank you sincerely.

And, of course, how could anything of this magnitude ever have been done without the timely execution and accurate scholarship of the contributing authors? They understood that the task at hand was to introduce educated readers (such as you) to new areas of interest in a very broad field, and without exception, they did a wonderful job. You will see throughout that their writing is clear and informative—just what material like this should be for the intelligent reader. To them, a sincere thank you and a job well done.

Finally, as always, none of this would have happened or been worth undertaking without my comrade in (almost all) ups and down and ins and outs, and my truest and best friend, Leni. Sara and Micah, versions 1.1 and 1.2, didn't hurt either.

—Neil J. Salkind
University of Kansas



Sometimes it is useful to know how large your zero is.

—Author unknown

ABILITY TESTS

Ability tests are assessment instruments designed to measure the capacity of individuals to perform particular physical or mental tasks. Ability tests were developed in the individual differences tradition of psychology and evolved from early tests of general intelligence. Most major ability tests assess a range of broad ability factors that are conceptually and empirically related to general intelligence (or *g*, also referred to as general cognitive ability). Ability tests are frequently used in settings such as schools, military organizations, business and industry, hospitals and rehabilitation centers, and private practice. Several ability tests with strong evidence of reliability and validity are currently available and are commonly used for purposes such as educational screening or diagnosis, personnel selection and classification, neuropsychological assessment, and career guidance and counseling.

Historical Overview

The first successful “mental test,” predecessor to all subsequent tests of characteristics of individual differences characteristics (including ability), is generally

considered to be the intelligence test developed by French psychologist Alfred Binet and his associate, Théodore Simon. First published in 1905, the Binet-Simon Intelligence Scale was designed to identify children presumably unable to benefit from regular classroom instruction by measuring their ability to judge, understand, and reason. The test was found to be an effective predictor of scholastic achievement. The success of the Binet-Simon scales and of later measures, such as Lewis M. Terman’s Stanford-Binet Intelligence Scale (published in 1916), led the emerging testing industry to focus on the further development of intelligence measures. Many of these early intelligence tests actually measured a range of different abilities.

At the outset of World War I, leading psychologists in the intelligence testing movement began attending to the problem of selecting and classifying recruits for the United States military. These efforts resulted in the development of group-administered intelligence tests such as the Army Alpha and Beta. The practical usefulness of these assessments and the efficiency with which they could be administered to large numbers of people led to the widespread use of tests and also to intensified research on specific areas of ability relevant to success in a variety of contexts. During the 1920s and 1930s, this shift from measures of general

intelligence to measures of specific abilities was accompanied by the development of a statistical technique called factor analysis. By identifying underlying factors on the basis of patterns of intercorrelations among a large number of variables, factor analysis made it possible to demonstrate that specific abilities (e.g., reading speed, reaction time) are indicators of broad areas of ability (e.g., broad visual perception, broad cognitive speediness) and that these broad abilities are somewhat independent of *g*.

Largely on the basis of evidence obtained from early factor analytic studies, two opposing theoretical approaches to understanding the ability domain emerged. The London school, led by Charles Spearman, emphasized *g* as the single most important ability. In contrast, a group of American scientists, led by Truman Kelley and Louis L. Thurstone, identified several relatively independent, broad ability factors. A classic study of mechanical ability, led by D. G. Paterson, provided early empirical evidence to support the claim that general areas of ability other than *g* accounted for significant variance in practical outcomes, such as job performance.

With World War II came the demand for follow-up efforts to the work conducted in the 1920s and 1930s. During the 1940s and 1950s, general multiple-ability test batteries, such as the Differential Aptitude Tests (DAT) and General Aptitude Test Battery (GATB), among many others, were developed and used frequently in subsequent decades. During the 1980s, controversy erupted over the question of fair use of the GATB, which was developed by the United States Employment Service, with prospective employees from racial and ethnic minorities. This controversy led to its suspension from use pending further study. A variety of alternative test batteries (three of which are reviewed below) measuring multiple areas of ability are available and in use today.

Definition and Dimensions of Ability

The construct of ability, as defined above, refers to the power of an individual to perform a specified act or task. Abilities are generally assumed to be fairly stable, to have a biological basis, and to be both learned

and innate. Ability may be differentiated from related constructs, such as *achievement*, which is defined as the level of knowledge or skill that has already been attained in an endeavor; *aptitude*, which is defined as the capacity to develop particular knowledge or skills in the future; and *intelligence*, which is typically defined as a general, higher-order ability relevant to tasks that have cognitive demands. These constructs clearly are related, and in practice the terms are sometimes used interchangeably. To further complicate the matter, tests of abilities or intelligence technically measure achievement and usually are used to infer aptitude. In the context of assessment, a general rule of thumb is as follows: achievement tests typically are designed to measure knowledge of a specified content area that has been explicitly taught; ability tests typically are designed to measure current performance in a particular content area or, when composed of a battery of subtests, across multiple content areas; and intelligence tests typically are designed to measure general cognitive ability. All three are used to infer aptitude, although the terms *aptitude tests* and *multiaptitude tests* usually refer to tests of ability.

There are a variety of theoretical approaches to understanding human ability, but the view most commonly held by scholars is that the domain of abilities can be represented using a hierarchical structure. For example, the Cattell-Horn-Carroll (CHC) theory of cognitive abilities, supported by one of the most comprehensive factor-analytic investigations of abilities in history, posits a three-level hierarchy. Like many hierarchical theories of abilities, CHC theory places *g* at the highest level, followed by several broad factors at the intermediate level of specificity and, in turn, by a large number of more narrowly defined, specific abilities at the lowest level (in the CHC model, these are approximately 70 in number). The 10 broad ability factors in the second level of the CHC model are similar to those posited by other hierarchical ability models and are as follows:

- *Fluid intelligence*: The ability to reason and solve problems involving novel information or procedures using processes that are not learned or culture bound

- *Crystallized intelligence*: The ability to communicate and reason using previously learned information and procedures
- *Quantitative knowledge*: The ability to manipulate numerical symbols and reason procedurally with quantitative information; includes mathematics achievement and knowledge
- *Short-term memory*: The ability to hold information in immediate awareness and effectively use it within seconds
- *Visual processing*: The ability to perceive, manipulate, analyze, and synthesize visual stimuli; includes visual memory and spatial relations
- *Auditory processing*: The ability to perceive, discriminate, analyze, and synthesize patterns in auditory stimuli; includes phonetic coding, memory for sound patterns, and ability to discriminate tones
- *Processing speed*: The speed with which information is attended to and processed; involves rapid, automatic cognitive processing
- *Long-term retrieval*: The ability to store information in long-term memory and accurately retrieve it later
- *Reading/writing*: The ability to read and understand written material accurately and efficiently and to write in a clear and organized manner with proper grammar, punctuation, and spelling
- *Decision/reaction time or speed*: The quickness with which problems of moderate difficulty are accurately encoded and mentally manipulated; includes simple reaction time and semantic processing speed

The preceding list of broad abilities in the CHC model is generally representative of the ability factors traditionally targeted for measurement by multiaptitude test batteries, although some differences exist across competing models of ability and measures of the domain. For example, because the CHC model targets cognitive ability, some abilities that are not traditionally considered cognitive in nature (e.g., psychomotor dexterity) and that are integral parts of other models of ability are excluded from the list. Also, some scholars have called for an expanded view of abilities that may include, for example, emotional intelligence, social intelligence, situational judgment, and other areas of human performance not typically included in traditional theoretical models of the ability domain.

Assumptions of Ability Tests

Although the specific features of various theoretical models underlying particular ability tests may differ, most major ability tests assume the following: (a) There are multiple abilities that can be reliably and validly measured using a single, wide-range test or battery; (b) there are differences between people in terms of level of performance in each area of ability; (c) there are differences within people in terms of level of performance across different areas of ability; (d) differences between a person's level of abilities relative to a normative group, and differences within a person's pattern of ability scores, predict real-world outcomes (e.g., academic and occupational performance); and thus (e) scores from ability tests offer useful information in settings where decisions related to education, employment, and rehabilitation are made. It should also be noted that ability tests measure *maximal* performance; some have proposed that *typical* performance may better predict some real-world outcomes.

Examples of Ability Tests

Several multiaptitude test batteries are currently available. Users are encouraged to select an instrument according to such criteria as the evidence for reliability and validity; the appropriateness of the normative samples used to standardize the test; the ease with which the test can be obtained, administered, and scored; the extent to which scale scores provide clear, unambiguous results; and the extent to which proposed applications of the test coincide with the needs of the user. The following are brief descriptions of three commonly used multiaptitude test batteries.

Armed Services Vocational Aptitude Battery

The Armed Services Vocational Aptitude Battery (ASVAB; U.S. Department of Defense) is best known for its use in military selection and classification and for its inclusion as part of a comprehensive career exploration program for high school and college students. The ASVAB consists of the following subtests, each separately timed: General Science, Arithmetic Reasoning, Word Knowledge, Paragraph

Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, Mathematical Knowledge, Mechanical Comprehension, and Electronics Information. Factor analytic evidence suggests that the ASVAB measures general cognitive ability, verbal-math ability, clerical speed, and technical knowledge. The ASVAB was developed using impressive norms, but its usefulness for differential prediction has been questioned.

Occupational Information Network Ability Profiler

The Occupational Information Network (O*NET) Ability Profiler (published by the U.S. Department of Labor) is a component of the O*NET Career Exploration Tools system. The O*NET Ability Profiler is an updated version of the GATB and is available in paper-and-pencil format with optional apparatus subtests. The O*NET Ability Profiler consists of 11 subtests that measure nine job-related abilities: verbal ability, arithmetic reasoning, computation, spatial ability, form perception, clerical perception, motor coordination, manual dexterity, and finger dexterity. A major strength of the battery is that it generates a computer-generated score report that can be linked to a wealth of occupational data in the O*NET database, allowing individuals to, for example, compare their pattern of abilities to those required by different occupations.

Differential Aptitude Tests

The DAT, published by the Psychological Corporation, was designed primarily for educational and career guidance of individuals in middle school, high school, and adulthood. The DAT provides scores for the following eight areas of ability: verbal reasoning, numerical ability, abstract reasoning, perceptual speed and accuracy, mechanical reasoning, space relations, spelling, and language usage. The DAT scores are computed using very good norm groups, and efforts to address test fairness have been thorough. Evidence supporting the reliability of the DAT scores is strong, although relatively little evidence is available to assess the validity of the scale scores for predicting outcomes other than academic achievement.

Each of the preceding test batteries has garnered evidence of reliability and validity, and this evidence is reviewed in the user's manuals. As indicated in their descriptions, the ASVAB, O*NET Ability Profiler, and DAT typically are used for educational and vocational counseling and personnel selection. Other purposes for testing abilities, such as neuropsychological evaluation or educational diagnosis, naturally require tests with validity evidence supporting their use for those purposes.

Gender and Ethnicity in Ability Testing

As is the case with any test, users must be sensitive to the ways in which personal or group characteristics such as age, gender, ethnicity, linguistic background, disability status, and educational and work history may influence performance on ability tests. Although the weight of evidence suggests no difference between females and males in general cognitive ability, consistent differences have been found favoring females on tests of some verbal abilities and males on tests of some visual-spatial tasks. Evidence also suggests that scores on quantitative abilities tend to favor females in the early years of school and males from adolescence onward. Most scholars suggest that biological and social factors work in tandem to produce such differences.

Some differences between ethnic groups have been found in scores on tests of ability. This finding contributed to the previously noted controversy over the GATB, for example. The crux of the controversy was that some minority groups tended to score lower than the majority on some GATB subscales and, since the U.S. Department of Labor had suggested that states use the test as part of a new employment selection system, members of these groups were adversely impacted. However, scores on the GATB (and other tests of ability) tended to predict educational and occupational outcomes equally well, regardless of ethnicity. Eventually the use of within-group norms in employee selection was proposed, but this suggestion was also controversial, and as mentioned earlier, the GATB was eventually suspended from use. Because

many tests measure abilities that are influenced by education and training, users must take into account the quality of the respondent's educational background. This is particularly important when interpreting scores of ethnically diverse respondents because minority groups are more likely than members of the majority to be socioeconomically disadvantaged, which in turn is related to poorer school systems and fewer educational opportunities that might improve test performance. Of course, it bears remembering that within-group differences are larger than between-group differences and that meaningful generalizations from the group to the individual can never be made responsibly without additional information.

Conclusion

Ability tests allow users to identify current and potential performance strengths for individuals, information that is useful for a wide range of purposes in a variety of contexts. Many ability tests are laden with positive features that likely contribute to their widespread use, such as ease of administration and scoring, strong psychometric evidence, and the provision of a large amount of meaningful information in a relatively brief period of time. With recent advances in technology, ability testing has become increasingly automated, and in the future, computer-administered testing will continue to improve convenience of use and also will allow users to customize their batteries to better suit their specific needs. When combined with other sources of information, the potential benefits of ability tests to individuals and society are substantial.

—Bryan J. Dik

See also Iowa Tests of Basic Skills; Reliability Theory; Validity Theory

Further Reading

- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Thurstone, L. L. (1938). *Primary mental abilities* (Psychometric Monograph No. 1). Chicago: University of Chicago Press.

U.S. Department of Labor. (2002). *Ability Profiler: Administration manual*. Washington, DC: U.S. Government Printing Office.

Armed Services Vocational Aptitude Battery: <http://www.asvabprogram.com/>

Cattel-Horn-Carroll Human Cognitive Abilities Project: www.iapsych.com/chchca.htm

Differential Aptitudes Tests: www.psychcorpcenter.com

O*NET Ability Profiler: www.onetcenter.org

ABSTRACTS

An abstract is a brief, concise, accurate, and generally nonevaluative summary of a work such as a journal article, a presentation, or a book. The length of an abstract varies but is typically a paragraph and never more than a page. An abstract for a periodical source often appears at the top of the article, underneath the title. For prepared manuscripts, an abstract is presented by itself on a single page that follows the title page. Abstracts are often collected in volumes and presented in either print or electronic format to provide potential readers of scholarly work with a quick and time-saving overview of the main document.

There are two common forms of abstracts. A *descriptive* abstract is often written prior to the completion of a specific work. Therefore, it may not provide results, conclusions, or recommendations. This type of abstract may be submitted to a local or a national conference, for instance, as a summary of one's planned presentation. A descriptive abstract may simply contain the problem and methods with a brief section on expected outcome. In contrast, an *informative* abstract is written following the completion of a specific work and summarizes the entire content of the original document. It commonly consists of an overview of the following four sections: (a) problem, (b) methodology, (c) results, and (d) conclusion. This type of abstract provides a condensed version of the original work so that a reader can choose whether to review the entire piece.

Abstract writing is an acquired skill that may be strengthened with continued practice. To present an effective abstract, the writer should follow the organization of the original article closely. In abstracts of scientific papers, the reader is first presented with information about the topic or central issue(s) in the main document. A statement about the study's objectives or the tested hypotheses may be provided. Second, the reader is educated about the methods used to approach the main topic. For example, the abstract may provide information relevant to the number of participants enrolled in a given study or the assessment strategies used to examine the stated hypotheses. Third, there is a brief description of the study's findings and conclusions. The reader is presented with an explanation of the significance and possible implications of the obtained results. Fourth, there is some reference to the recommendations. In summary, a well-written, self-contained abstract presents a capsule description of the original article, without adding new information, in language that is understandable to a wide audience.

—Marjan Ghahramanlou-Holloway

See also American Doctoral Dissertations; American Psychological Association

Further Reading

- Hartley, J. (1998). *An evaluation of structured abstracts in journals published by the British Psychological Society*. Retrieved from <http://cogprints.org/587/00/199801001.html>
- Kamler, B., & Thomson, P. (2004). Driven to abstraction: Doctoral supervision and writing pedagogies. *Teaching in Higher Education*, 9(2), 195–209.
- Ono, H., Phillips, K. A., & Leneman, M. (1996). Content of an abstract: De jure and de facto. *American Psychologist*, 51(12), 1338–1340.
- Pierson, D. J. (2004). How to write an abstract that will be accepted for presentation at a national meeting. *Respiratory Care*, 49(10), 1206–1212.

ACCEPTANCE SAMPLING

Acceptance sampling is a procedure used for product acceptance or rejection and is based on inspecting

only a sampled number of units from the total number produced. In many situations in which the inspection is destructive, such as testing flashbulbs, it is not feasible to inspect all the bulbs produced. Acceptance sampling can be performed during incoming inspection of raw materials or components, in various phases of in-process operations, or during final inspection. Such a procedure may be applied to cases in which inspection is by attributes or by variables.

Acceptance sampling is a scheme that determines whether a batch or lot of product items should be accepted or rejected. It does not control or improve the quality level of the process. Although acceptance sampling offers some advantages associated with its feasibility in destructive testing, its economies in inspection cost, and its usefulness in improving quality, it also poses some risks because an entire batch may be rejected based on inspection of a few items.

Risks in Acceptance Sampling

Two types of risks are inherent in acceptance sampling plans: *producer's risk* and *consumer's risk*.

Producer's Risk

This is the risk (denoted by α) associated with rejecting a lot that is of "good" quality, and a numerical definition of a "good" lot is prescribed by the *acceptable quality level* (AQL). The AQL may be viewed as the maximum proportion of nonconforming items in a batch that can be considered satisfactory as a process average. Thus, the interpretation of the statement that the producer's risk is 5% for an AQL of 0.03 is as follows: Batches that are 3% nonconforming are considered satisfactory, and it is desirable that such batches, or those that are better, not be rejected more than 5% of the time.

Consumer's Risk

This is the risk (denoted by β) associated with accepting a "poor" lot. A numerical definition of a poor lot is indicated by the limiting quality level (LQL), also referred to as the lot tolerance percent defective. The statement that the consumer's risk is

10% for an LQL of 0.09 means that batches that are 9% or more nonconforming are considered poor. Consequently, such batches should be accepted no more than 10% of the time by the selected acceptance sampling plan.

Acceptance Sampling by Attributes and Variables

In attribute acceptance sampling, a product is classified as nonconforming or unacceptable if it contains one or more nonconformities or defects. For example, a hair dryer may be nonconforming if the speed control switch does not operate at each of the settings.

Alternatively, acceptance sampling may be used in the context of product-related variables, which can be measured and on the basis of which a decision about the product can be made. An example is the monitoring of the weight of cereal boxes. Suppose a minimum acceptable weight of 12 ounces is specified. By the selection of reasonable protection levels associated with errors that could be made in decision making using acceptance sampling, an *acceptance limit* is found. For sampled cereal boxes, sample statistics such as the mean and the standard deviation may be calculated. Using these measures, if the calculated sample mean is less than the acceptance limit, the batch of boxes could be rejected.

Parameters of an Attribute Sampling Plan

In the simplest of the attribute sampling plans, namely, a single sampling plan, a decision is made on a batch or lot on the basis of information from a single sample. There are two parameters for such a plan, the *sample size* (n) and the *acceptance number* (c). These parameters are chosen on the basis of acceptable levels of the producer's (α) and consumer's (β) risks and values of AQL and LQL. Standardized procedures exist whereby, on the basis of on chosen levels of the above-mentioned parameters, the values of n and c may be determined.

Suppose, for example, the batch size (N) for a product is 2,000. On the basis of a producer's risk (α) of 5%, AQL of 2%, a consumer's risk (β) of 10%, and

an LQL of 8%, the sample size (n) and acceptance number (c) are determined to be, say, 50 and 3, respectively, for the single sampling acceptance plan. The plan operates as follows. A sample of 50 units is randomly selected from the batch of 2,000 units, and the number of nonconforming units is found. If the number of nonconforming units is less than or equal to the acceptance number, which in this example is 3, the batch is accepted. Otherwise, the batch is rejected.

—Amitava Mitra

See also Hypergeometric Distribution; Sample; Sampling Error

Further Reading

American National Standards Institute (ANSI)/American Society for Quality (ASQ). (2003). *Sampling procedures and tables for inspection by attributes*, ANSI/ASQ Z1.4. Milwaukee, WI: ASQ Quality Press.

American National Standards Institute (ANSI)/American Society for Quality (ASQ). (2003). *Sampling procedures and tables for inspection by variables for percent nonconforming*, ANSI/ASQ Z1.9. Milwaukee, WI: ASQ Quality Press.

Mitra, A. (2005). *Fundamentals of quality control and improvement* (2nd ed.). Mason, OH: Brooks/Cole.

ACHIEVEMENT TESTS

Any test that is designed to measure student learning in the context of an educational or training program can be called an achievement test. An achievement test comprises one to many test items. Each test item can be scored dichotomously (right or wrong) or with a rating scale, on which degrees of performance are determined by a judge, called a rater. Achievement test items are usually distinguished by the kind of response they generate: selected or constructed. The selected response item is often referred to as multiple choice because the test respondent chooses among the choices offered. The constructed response item requires that the respondent generate a written or oral response or a response in the form of a product or process. There is considerable variety in selected and constructed response test items.

The publication *Standards for Educational and Psychological Testing* provides definitions relevant to achievement tests. This publication also offers many guidelines for the development and validation of achievement tests. Another useful, comprehensive reference about the development and validation of achievement tests is the *Handbook of Test Development*.

The length of an achievement test varies according to many factors, all of which relate to validity and the intended purpose of the test. One of the most important factors is reliability. Longer tests tend to yield more reliable scores. If an achievement test represents a domain of identifiable knowledge and skills, then it should be a representative sample from this domain. Reliability and adequate sampling of content are two major types of evidence that support the validity of an achievement test score interpretation or use.

Another useful distinction is the difference between a *test* and a *quiz*. A quiz is shorter than a test and measures only several student learning objectives, whereas a test is longer than a quiz and measures many student learning objectives. Both test and quiz, as distinguished here, constitute a measure of student achievement, but the distinction is in the amount of coverage of the domain or content to be learned. Thus, any quiz is also an achievement test in this broader sense.

The term *assessment* is often used synonymously with the term *achievement test*. Strictly speaking, the two are not the same. Assessment is a judgment, usually by a teacher, about how well a student has learned and what a student needs to learn. An assessment should be based on valid information, which includes results of achievement tests and other information collected during a semester or school year or in a training program. Thus, an achievement test that is used for an assessment purpose may be given the name *assessment* because the test information is used for an assessment of student learning.

What Is Achievement?

Achievement is generally considered change in cognitive behavior that we attribute to learning, which can occur both within and outside a planned learning experience, such as a course, class, or training. What

students learn can be thought of as existing in one of two related domains. Each domain has a large collection of test items that represent it.

The first domain, which is more familiar to educators, consists of knowledge and skills. We have many examples, including reading, mathematics, and social studies. Generally, states and school districts have content standards that explicitly define each domain. One type of achievement test is intended to be a representative sample from a domain of knowledge and skills. This kind of achievement test is generally created in a selected-response format because of the great efficiency of this format when compared to a constructed-response format. One of the major shortcomings of the latter format is that it does not yield an adequate sample from the intended domain of tasks that represent achievement.

The second domain, which is less familiar to educators, consists of many complex tasks that represent an ability. For example, writing is an ability. The domain of tasks that represent this ability may include making a report based on some experience; writing a creative story; writing a memorandum, e-mail, or letter; writing an invitation to a social event; and writing a critique, among many others. Many educators and psychologists have suggested that each ability is complex in nature. Each ability is learned slowly and unevenly over a lifetime. Each ability requires the application of knowledge and skills in unique ways to situations that we commonly encounter. The abilities that we learn in school are reading, writing, speaking, listening, mathematical and scientific problem solving, critical thinking, and creative thinking. Most states have adopted content standards that contain learning objectives that describe the kinds of student behaviors that can be tested. The testing format that seems most appropriate for this type of content is performance based. The scoring of these performance-based tests can be done subjectively, using trained, experienced raters and rating scales, which are referred to as *rubrics*, or the scoring can be done objectively, as we often see in mathematics, where there is likely to be a correct answer or a step-by-step algorithm to apply to reach a correct answer.

The scores from these achievement tests can be reported as the number of correct answers (raw score); as a percentage of the total number of items on the test; as a percentile rank or stanine; or in some derived scale, such as grade equivalent, standard score, or normal-curve equivalent. The choice of a scale to report a person's score depends on many factors related to the purpose of the test. *Norm-referenced* interpretations of test scores value the rank of scores in a set of test scores. Knowing how students rank in achievement is useful for some test purposes. *Criterion-referenced* interpretations are more concerned with how much a student has learned, usually relative to a standard, called a *cut score*. Another term for criterion-referenced is *domain-referenced* because the score a student obtains (e.g., 75% correct) refers to the degree of learning that has occurred relative to a domain of knowledge and skills. The terms *criterion-referenced test* and *norm-referenced test* are commonly used, but like the term *assessment*, they are not used accurately. Any test can yield a norm-referenced interpretation by using a norm-referenced test score scale. Many achievement tests lend themselves to criterion-referenced or domain-referenced interpretations due to the way they were designed. Thus, we use the terms *norm-referenced test* and *criterion-referenced test* to refer to the type of interpretation we desire from the test, but strictly speaking, these are not types of tests.

The Distinction Between Achievement Tests and Intelligence Tests

Psychologists and others have spent more than a hundred years studying human intelligence. Depending on one's experiences and background, this field has been controversial, but some ideas have survived this controversy. Intelligence and achievement are often considered as being on a continuum in the cognitive domain. Achievement is generally viewed as something that changes with experience and instruction. Achievement tests can detect these changes, which we know as student learning. Intelligence is generally viewed as a group of complex abilities that

are less resistant to change. These abilities are verbal, quantitative, and analytical. Some standardized tests are indicators of intelligence, and other tests are indicators of achievement. Sometimes the distinction is not clear. Validation is a process whereby the truth of what a test measures is studied and the claim for what it measures is supported by reasoning and evidence.

Purpose

An achievement test can also be distinguished by its purpose. A major difference in purpose distinguishes a classroom achievement test and the standardized achievement test. The design of each type of achievement test and the uses of its test scores will vary according to its purpose.

Classroom Achievement Tests

A classroom achievement test is a category that seems focused on two central purposes: *formative* and *summative*. During instruction, formatively used achievement tests inform both student and teacher about the extent of learning. These achievement tests do not count toward a student grade. Instead, formatively used achievement tests guide the student to improve learning. Summatively used tests are used for some important purpose, such as grading. During a grading period, usually 9 weeks, a test can be used as part of the criteria for a student grade. A test score, by itself, is never recommended by evaluation specialists as the sole indicator of a grade. A test result is only one important piece of information about the extent and quality of student learning.

The validity of formative and summative tests depends on several factors. First, the content must be clearly stated to students. Second, factors that prevent learning should be removed or minimized. Third, instruction should be aligned to this content. Fourth, the achievement test should also be aligned to this content. Fifth, test design and administration can be modified to remove factors that may invalidate test performance. Finally, students should have additional opportunities to learn what they have not learned.

Standardized Achievement Tests

Standardized achievement tests are numerous. Two comprehensive references on standardized achievement testing are the *Sixteenth Mental Measurement Yearbook* and *Tests in Print*. A useful Web address for these and similar publications is <http://www.unl.edu/buros/>.

A major distinction among standardized tests is whether they are intended to help assess student learning in an educational program or to help determine who passes or fails for a promotion, graduation, certification, or licensing decision. Achievement tests that have significant consequences for the test taker or the public are considered *high-stakes* achievement tests.

Standardized achievement tests have common characteristics. The population that is to take this test is well described. The content of the test is clearly specified. Items are systematically developed and validated. The test is designed to maximize information that will be validly interpreted and used by its recipients. The conditions for administration and scoring are uniform. If multiple test forms exist, these test forms are usually equated so that interpretations are consistent from test form to test form. Standards are set validly using procedures that are widely known and accepted as producing valid results. The interpretation of test scores is consistent with the intended purpose of the test. Thus, the term *standardized* refers to these aspects of test design, test development, administration, scoring, and reporting.

To ensure that the test scores are validly interpreted and used, all standardized achievement tests are subject to validation. The process of validation is a responsibility of the test developer and the sponsor of the testing program. The *Standards for Educational and Psychological Testing* are very clear about the conditions and evidence needed for validation. A technical report or test manual contains the argument and evidence supporting each intended test score interpretation or use. If subscores are used, each subscore should also be validated. Test scores should never be used for purposes that have not been validated.

A major distinction among standardized achievement tests is whether a test has been aligned to a set of content standards *specifically* or *generally*.

Specifically aligned tests provide the public with assurance that what content is mandated in a state or a school district is represented on the aligned test. An alignment ensures that curriculum, instruction, and test all correspond to each another. A good example of a set of national content standards is the one developed in 1991 by the National Council of Teachers of Mathematics. It is reasonable to expect that a state's or school district's aligned test scores will also be based on national standards and that test scores from a state- or district-aligned test will correlate highly with scores from one of several generally aligned achievement tests, such as the Stanford Achievement Test, the Iowa Tests of Basic Skills, or the TerraNova.

All standardized achievement tests can be used for various purposes. However, each purpose should be validated. For instance, some standardized achievement tests can be used diagnostically to identify a student's strengths and weaknesses. The same tests can also be used to identify strengths and weaknesses in a curriculum or instructional program for a school or a school district. This kind of analysis can even be done at the state, national, or international level. In some instances, a standardized achievement test can be used for making pass-fail decisions for promotion to a higher grade or for graduation. Validation should focus on the legitimacy of using any test score for any purpose.

Summary

Any test that is intended by its developer to reflect the amount of learning that has occurred in the past can be considered an achievement test. Assessment is a term often mistakenly used for a test. Assessment involves the use of information about student learning to reach a conclusion about what a student knows and can do and what the student needs to learn. The content of achievement tests can be considered in two ways: as a domain of knowledge and skills and as a domain of complex tasks. To complete a complex task, a student has to have knowledge and skills and be able to apply each in a complex way to perform the task. Achievement tests are of two major types: (a) tests used in the classroom for formative or summative assessment and (b) standardized tests, which serve

many purposes, including assessment. Any use of a test should be validated.

—Thomas Haladyna

See also Iowa Tests of Basic Skills; Iowa Tests of Educational Development

Further Reading

- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.
- Kane, M. T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher*, 22, 12–23.
- Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (2002). *Tests in print VI*. Lincoln: University of Nebraska Press.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurement yearbook*. Lincoln: University of Nebraska Press.

Buros Center for Testing: <http://www.unl.edu/buros/>

ACTIVE LIFE EXPECTANCY

Active life expectancy (ALE) refers to the number of years of life one can be expected to live without a disability. ALE answers the question, Of the remaining years of life for this cohort of persons, what proportion is expected to be spent disability-free? As such, it is used to evaluate the quality of life rather than just the quantity. The term was first introduced by Katz and colleagues in 1983, although others had described similar concepts in the mid-1960s and early 1970s. ALE is a summary measure of population health, is based on aggregate statistics, and is used in demography and epidemiology to measure and compare the health and functional status of national populations.

ALE is usually expressed as a component of total life expectancy at a certain age. In the measurement of ALE, disability is usually defined as difficulty in performing one or more activities of daily living (ADLs), which include eating, dressing, bathing, toileting, walking, and transferring to a bed or chair. There are two closely related concepts, disability-free life expectancy (DFLE) and healthy, or health-adjusted, life expectancy (HALE). While ALE uses the presence of a limitation in any one of the activities of daily living as an endpoint, disability-free life expectancy uses the presence of limitations in either ADLs or instrumental activities of daily living (IADLs), and HALE uses a measure of general health status, or good health versus poor health.

ALE can be computed in one of two ways, by using multistate life tables or by the Sullivan method. The Sullivan method requires a period life table for the population and observed age-specific prevalence of disability in the population (π_i). Using standard life table notation, at each age i , the number of person years lived in that age interval (L_i) is multiplied by the proportion of people at that age who are not disabled ($1 - \pi_i$). ALE is the total of all person years lived above an age i divided by the number of survivors to age i , l_i . In contrast, the multistate life table method requires age-specific transition probabilities. These transition probabilities, usually derived from longitudinal survey data, give the rates of moving from health to disability (and back), as well as the risk of death for both the healthy and disabled states at each age. Survival models incorporate these transition probabilities to estimate the time expected to be spent in the two states, active and disabled. The advantages of the multistate method include the ability to model recovery and the ability to include covariates in the models.

This method was used in 2005 by Reynolds and colleagues to examine the effect of obesity on ALE. They found that being obese at age 70 has virtually no effect on the total number of years of life remaining but reduces ALE by 2.4 years for women and 1.5 years for men.

—Christine L. Himes

See also Longitudinal/Repeated Measures Data; Survival Analysis

Further Reading

- Katz, S., Branch, L. G., Branson, M. H., Papsidero, J. A., Beck, J. C., & Greer D. S. (1983). Active life expectancy. *New England Journal of Medicine*, 309, 1218–1224.
- Reynolds, S. L., Saito, Y., & Crimmins, E. M. (2005). The impact of obesity on active life expectancy in older American men and women. *Gerontologist*, 45, 438–444.
- Sanders, B. (1964). Measuring community health level. *American Journal of Public Health*, 54, 1063–1070.
- Schoen, R. (1988). *Modeling multigroup populations*. New York: Plenum.
- Sullivan, D. F. (1971). A single index of mortality and morbidity. *HSMHA Health Reports*, 86, 347–354.

ADAPTIVE SAMPLING DESIGN

In adaptive sampling, information gained during the sampling process is used to modify, or adapt, how the subsequent sample units are selected. Traditionally, the selection procedure is defined prior to sampling. For example, a sample scheme may be to select n units at random from the population, as in simple random sampling. In adaptive sampling, the select procedure may change during the survey.

In biology, plant or animal populations are often spatially aggregated, and when the population is rare, sampling can be challenging. Adaptive sampling can be used in this situation. An initial sample of plants may be undertaken within quadrats placed randomly in the area. Additional sampling is undertaken around the quadrats where plants were found. The final sample combines information from the initial sample and from the additional sampling. In this adaptive design, survey effort was concentrated in the localities where plants were found. The information from the initial sample was used to “adapt” the sample because additional sampling was directed to the localities where plants were known to be. In this regard, adaptive sampling is considered more informative than traditional sampling.

There are many forms of adaptive sampling; the one described above could be called adaptive cluster sampling. Another adaptive design is to adaptively allocate extra survey effort in stratified sampling. In traditional stratified sampling, the population is

divided into homogeneous groups or regions, called strata. Survey effort is allocated among strata according to some criterion, usually estimates of the within-stratum variance or mean. If there is no information on these strata statistics or if the information is poor, adaptive stratified sampling can be used. After an initial survey of the strata, estimates of stratum variance or mean are used to decide on allocation of additional survey effort. Usually this additional survey effort is allocated to the strata with the highest variances. Information on the strata gained during the survey is used to adapt the final allocation of survey effort.

—Jennifer Ann Brown

See also Sample; Sampling Error

Further Reading

- Smith, D. R., Brown, J. A., & Lo, N. C. H. (2004). Applications of adaptive sampling to biological populations. In W. L. Thompson (Ed.), *Sampling rare or elusive species* (pp. 77–122). Washington, DC: Island Press.
- Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050–1059.
- Thompson, S. K., & Seber, G. A. F. (1996). *Adaptive sampling*. New York: Wiley.

ADJECTIVE CHECKLIST

The Adjective Checklist (ACL) is a measure of children’s attitudes that utilizes a checklist format first employed by Harrison Gough. The ACL has been employed in more than 30 studies to assess children’s attitudes toward persons from potentially stigmatized groups, with a focus on peers with mental retardation. Other studies have examined attitudes toward children with visual impairments, autism, obesity, cancer, and physical disabilities, as well as toward tobacco users.

The ACL was developed to assess the *cognitive* component of children’s attitudes (opinions and beliefs about a person), one of three components that make up attitudes (together with the *affective* component, i.e., emotions and feelings about a person, and the *behavioral*

intentions component, i.e., intentions to interact with a person). It uses an open-ended format that allows children to select from a provided list as many positive and negative adjectives as they wish to select to describe a specific person (known as a *target*). The open-ended approach of the ACL does not restrict children to making judgments that they may not ordinarily make, the way a forced choice format might. That is, the ACL mirrors the behavior of children in classroom settings where children express their opinions or beliefs about a peer by using common descriptors such as “smart,” “mean,” “friendly,” and so on.

The ACL was developed by asking large samples of children in Grades 1 through 6 to identify terms they would use to describe a person they liked and a person they did not like. Those terms that were mentioned most often were compiled into a list, and new samples of children were asked to judge each term as a “good” thing or a “bad” thing to say about someone. As a result, 34 adjectives were identified that describe a person’s affective feelings, physical appearance, academic behavior, and social behavior. Within these broad categories, the ACL includes equal numbers of positive and negative descriptors. Factor analysis of ACL responses from more than 2,000 elementary school children revealed three distinct factors: positive (P factor, e.g., “proud,” “happy”), negative (N factor, e.g., “careless,” “ugly”), and negative affect (NA factor, e.g., “lonely,” “ashamed”).

The ACL can be administered to children individually or in groups by asking the children to use the checklist to describe a particular target. The target may be either a hypothetical student depicted in a videotaped vignette, a photograph, a verbal description, or a real individual. In each instance, the target is presented, and then the children are asked to describe the target using as few or as many words from the list as they would like. There are two methods for scoring the ACL. The first method involves summing up a child’s selection of adjectives on each of the three factors noted above. The second method results in a composite score in which the number of negative adjectives chosen by a child is subtracted from the number of positive adjectives chosen, and a constant of 20 is added. In this method, the negative adjectives

would include all adjectives in the N factor and the NA factor (i.e., Total Score = P – N – NA + 20). A resulting score below 20 represents a negative attitude toward the target, and a score above 20 represents a positive attitude.

The ACL has good construct validity: Correlations with measures of behavioral intentions include Pearson *r* values of .76 with the Foley Scale and .67 with the Siperstein Activity Preference Scale, .35 with Selman’s Friendship Activity Scale, and .46 with the Shared Activities Questionnaire. Cronbach’s alpha has been reported to range from .67 to .91, with values of .83, .76, and .73 reported for the P, N, and NA factors, respectively.

—Gary N. Siperstein

See also Attitude Tests

Further Reading

- Bak, J. J., & Siperstein, G. N. (1987). Similarity as a factor effecting change in children’s attitudes toward mentally retarded peers. *American Journal of Mental Deficiency, 91*(5), 524–531.
- Bell, S. K., & Morgan, S. B. (2000). Children’s attitudes and behavioral intentions toward a peer presented as obese: Does a medical explanation for the obesity make a difference? *Journal of Pediatric Psychology, 25*(3), 137–145.
- Campbell, J. M., Ferguson, J. E., Herzinger, C. V., Jackson, J. N., & Marino, C. A. (2004). Combined descriptive and explanatory information improves peers’ perceptions of autism. *Research in Developmental Disabilities, 25*(4), 321–339.
- Castagano, K. S. (2001). Special Olympics unified sports: Changes in male athletes during a basketball season. *Adapted Physical Activity Quarterly, 18*(2), 193–206.
- Gough, H. G. (1952). *The Adjective Check List*. Palo Alto, CA: Consulting Psychologist Press.
- Graffi, S., & Minnes, P. M. (1988). Attitudes of primary school children toward the physical appearance and labels associated with Down syndrome. *American Journal of Mental Retardation, 93*(1), 28–35.
- Kury, S. P., Rodrigue, J. R., & Peri, M. G. (1998). Smokeless tobacco and cigarettes: Differential attitudes and behavioral intentions of young adolescents toward a hypothetical new peer. *Journal of Clinical Child Psychology, 27*(4), 415–422.
- Manetti, M., Schneider, B. H., & Siperstein, G. N. (2001). Social acceptance of children with mental retardation:

- Testing the contact hypothesis with an Italian sample. *International Journal of Behavioral Development*, 25, 279–286.
- Siperstein, G. N. (1980). *Instruments for measuring children's attitudes toward the handicapped*. Boston: University of Massachusetts.
- Siperstein, G. N., & Gottlieb, J. (1997). Physical stigma and academic performance as factors affecting children's first impressions of handicapped peers. *American Journal of Mental Deficiency*, 81, 455–462.
- Swaim, K. F., & Morgan, S. B. (2001). Children's attitudes and behavioral intentions toward a peer with autistic behaviors: Does a brief educational intervention have an effect? *Journal of Autism & Developmental Disorders*, 31(2), 195–205.

AGE NORMS

Age norms are used to represent typical performance or some aspect of development for children within a particular age group. Used as an indication of the average age at which certain behaviors are expected to occur, they provide a metric against which same-aged peers can be compared. Alternatively, they provide guidelines to determine where along a developmental continuum an individual's skill or behavior may fall. Depending on the measure of interest, these norms may be expressed in various ways.

The use of age norms assumes homogeneity of a group with respect to particular skills or behaviors. Because these can be expected to be normally distributed within a population, age norms can be computed on the basis of the average performance of the individuals within that population. For example, a vocabulary of 50 words is considered to be the norm for typically developing children between the ages of 12 and 18 months. Children whose vocabulary size falls within, above, or below this range may therefore be considered typical, precocious, or delayed, respectively. Age norms exist as well for certain physiological measures (e.g., the pitch of the voice) as well as developmental milestones (e.g., crawling or walking).

Age norms are also employed in characterizing the acquisition or emergence of certain skills. These norms assume an ordering of developmental stages

and are often used to characterize motor functions, aspects of speech and language acquisition, social behaviors, and so forth. Often, the emergence of behaviors is considered to be predicated on the acquisition of prerequisite skills, thus implying a fixed and orderly developmental sequence. This pattern would be typical of sensorimotor phenomena such as locomotion and manual dexterity. For example, the ability to stabilize the trunk using large muscle groups typically occurs at a certain age and precedes the development of movements necessary for more precise distal movements. By extension, failure to develop earlier skills would predict the delay, impairment, or absence of later-emerging skills.

Other behaviors may also appear along a developmental continuum. For example, starting as early as 1 year of age, norms exist for the production of classes of speech sounds (e.g., stop consonants vs. fricatives), individual sounds within those classes, the ways those sounds are used in words, syllable structure, and so on. Although motorically complex sounds often appear later than simpler ones, a fixed order does not necessarily apply. Age norms exist as well for the acquisition of grammatical structures and various parts of speech. Failure to acquire speech and language according to these norms is considered grounds for further evaluation or intervention.

Age norms are typically easy to understand, with respect to performance both at a particular age and over time. However, their usefulness is limited to certain types of developmental behaviors or skills. Moreover, although skills or milestones that fall at or within age norms may be considered to be normal or typical, interpretation is more problematic when performance lies outside those norms. As a result, measures are typically standardized so that (a) a child's performance can be characterized with respect to other children of the same age or grade and (b) a comparison of performance across different assessment instruments can be made for the same child.

—Carole E. Gelfer

See also Data Collection; Longitudinal/Repeated Measures Data

Further Reading

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Sherry, A., Henson, R. K., & Lewis, J. G. (2003). Evaluating the appropriateness of college-age norms for use with adolescents on the NEO Personality Inventory—Revised. *Assessment*, 10(1), 71–78.

The NEO Personality Inventory measures normal personality characteristics and has demonstrated appropriate score reliability and validity. **Age norms** are available for two groups of individuals, college-age individuals 17 to 20 years old and adults 21 and older. Often, personality instruments normed on older individuals have been used with adolescent populations. To examine the appropriateness of this decision, the current study explored the differences between an adolescent sample and a college-age sample on the 30 facets and the five domains of the NEO. Group differences on the facet and domain scales were analyzed using discriminant analysis. Results indicated that the adolescent and college groups differed on each of the five domains. As expected, the groups also scored differently when the aggregated domain-level variables were used as the outcome measures.

AKAIKE INFORMATION CRITERION

In statistical modeling, one of the main challenges is to select a suitable model from a candidate family to characterize the underlying data. Model selection criteria provide a useful tool in this regard. A selection criterion assesses whether a fitted model offers an optimal balance between goodness-of-fit and parsimony. Ideally, a criterion will identify candidate models that are either too simplistic to accommodate the data or unnecessarily complex.

The Akaike information criterion (AIC) was the first model selection criterion to gain widespread acceptance. AIC was introduced in 1973 by Hirotugu Akaike as an extension to the maximum likelihood

principle. Conventionally, maximum likelihood is applied to estimate the parameters of a model once the structure of the model has been specified. Akaike's seminal idea was to combine estimation and structural determination into a single procedure.

The minimum AIC procedure is employed as follows. Given a family of candidate models of various structures, each model is fit to the data via maximum likelihood. An AIC is computed based on each model fit. The fitted candidate model corresponding to the minimum value of AIC is then selected.

AIC serves as an estimator of Kullback's directed divergence between the generating, or "true," model (i.e., the model that presumably gave rise to the data) and a fitted candidate model. The directed divergence assesses the disparity or separation between two statistical models. Thus, when entertaining a family of fitted candidate models, by selecting the fitted model corresponding to the minimum value of AIC, one is hoping to identify the fitted model that is "closest" to the generating model.

Definition of AIC

Consider a candidate family of models denoted as $M_1, M_2, \dots, M_L$. Let θ_k ($k = 1, 2, \dots, L$) denote the parameter vector for model M_k , and let d_k denote the dimension of model M_k , that is, the number of functionally independent parameters in θ_k . Let $L(\theta_k | y)$ denote the likelihood for θ_k based on the data y , and let $\hat{\theta}_k$ denote the maximum likelihood estimate of θ_k . The AIC for model M_k is defined as

$$AIC_k = -2\log L(\hat{\theta}_k | y) + 2d_k.$$

The first term in AIC_k , $-2\log L(\hat{\theta}_k | y)$, is based on the empirical likelihood $L(\hat{\theta}_k | y)$. This term, called the goodness-of-fit term, will decrease as the conformity of the fitted model M_k to the data improves. The second term in AIC_k , called the penalty term, will increase in accordance with the complexity of the model M_k . Models that are too simplistic to accommodate the data are associated with large values of the goodness-of-fit term, whereas models that are unnecessarily complex are associated with large values of

the penalty term. In principle, the fitted candidate model corresponding to the minimum value of AIC should provide an optimal tradeoff between fidelity to the data and parsimony.

The Assumptions Underlying the Use of AIC

AIC is applicable in a broad array of modeling frameworks because its justification requires only conventional large-sample properties of maximum likelihood estimators. However, if the sample size n is small in relation to the model dimension d_k (e.g., $d_k \approx n/2$), AIC_k will be characterized by a large negative bias. As a result, AIC_k will tend to underestimate the directed divergence between the generating model and the fitted candidate model M_k . This underestimation is potentially problematic in applications in which the sample size is small relative to the dimensions of the larger models in the candidate family. In such settings, AIC may often select a larger model even though the model may be unnecessarily complex and provide a poor description of the underlying phenomenon. Small-sample variants of AIC have been developed to adjust for the negative bias of AIC. The most popular is the “corrected” AIC (AICc), which was first proposed in 1978 for the framework of normal linear regression by Nariaki Sugiura. A decade later, AICc was generalized, advanced, and popularized in a series of papers by Clifford M. Hurvich and Chih-Ling Tsai.

AIC can be used to compare nonnested models. AIC can also be used to compare models based on different probability distributions, such as normal versus Poisson. However, if the models in the candidate family are based on different distributions, all terms in each empirical likelihood must be retained when the values of AIC are evaluated. (If the models in the candidate family are based on the same distribution, terms in the empirical likelihood that do not depend on the data may be discarded in the AIC computations.) AIC cannot be used to compare models

based on different transformations of the response variable.

An Application

The following data set appears in Annette J. Dobson’s text *An Introduction to Generalized Linear Models* (2nd ed.), 2002, pp. 51–53. The response variable y_i consists of the number of deaths from coronary heart disease in a 5-year age group for men in the Hunter region of New South Wales, Australia, in 1991. Table 1 features the values of y_i , the age groups and the group indices i , the population sizes n_i , and the mortality rates per 100,000 men (i.e., $y_i / n_i \times 100,000$).

Figure 1 features a plot of the log of the mortality rate (per 100,000 men) versus the age group index i . Dobson notes that the plot is approximately linear. Thus, if $\mu_i = E[y_i]$, the plot might suggest that $\log(\mu_i / n_i)$ could be modeled as a linear function of the group index i . If the response y_i is regarded as a Poisson random variable, then a generalized linear model (GLM) of the following form might be postulated for the data:

$$M_1: \log \mu_i = \log n_i + \alpha + \beta_1 i, y_i \sim \text{Poisson}(\mu_i).$$

However, one could also argue that the plot exhibits slight curvature for the older age groups and that the

Table 1 Number of Deaths From Coronary Heart Disease and Population Sizes by 5-Year Age Groups for Men in the Hunter Region of New South Wales, Australia, 1991

Group index, i	Age group (years)	Number of deaths, y_i	Population size, n_i	Mortality rate per 100,000 men, $y_i/n_i \times 100,000$
1	30–34	1	17,742	5.6
2	35–39	5	16,554	30.2
3	40–44	5	16,059	31.1
4	45–49	12	13,083	91.7
5	50–54	25	10,784	231.8
6	55–59	38	9,645	394.0
7	60–64	54	10,706	504.4
8	65–69	65	9,933	654.4

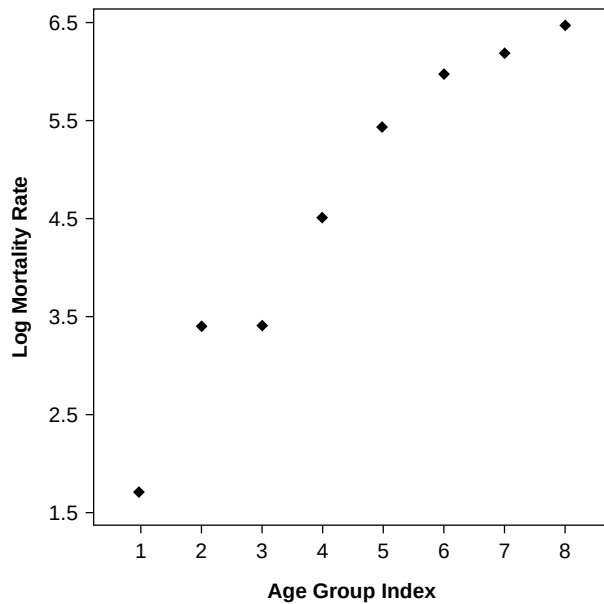


Figure 1 Log Mortality Rate (per 100,000 Men) Versus Age Index

mean structure of the model should account for this curvature. Following this reasoning, an alternative GLM might be postulated that describes $\log(\mu_i/n_i)$ as a quadratic function of i :

$$M_2: \log \mu_i = \log n_i + \alpha + \beta_1 i + \beta_2 i^2, y_i \sim \text{Poisson}(\mu_i).$$

AIC provides a tool for determining which model is more appropriate, M_1 or M_2 . If the GLM's M_1 and M_2 are fit to the data using PROC GENMOD in SAS (version 9.1), the empirical log-likelihood $\log L(\hat{\theta}_1 | y)$ is produced as part of the standard output. For model M_1 , we have $\theta_1 = (\alpha, \beta)'$, $d_1 = 2$, $\log L(\hat{\theta}_1 | y) = 539.0088$, and $\text{AIC}_1 = -1074.02$. For model M_2 , we have $\theta_2 = (\alpha, \beta_1, \beta_2)'$, $d_2 = 3$, $\log L(\hat{\theta}_2 | y) = 544.8068$, and $\text{AIC}_2 = -1083.61$. Thus, the minimum AIC procedure favors the quadratic model, M_2 .

For model M_2 , the Wald test based on the null hypothesis $\beta_2 = 0$ yields a p value of .0016. Thus, the Wald test further supports the inclusion of the quadratic term. (The statistics for marginal Wald tests are also produced as standard output for PROC GENMOD in SAS.)

We include the Wald p value for the test of $\beta_2 = 0$ merely for the sake of illustration. In general, the use of AIC should not be combined with hypothesis

testing, because each tool is formulated according to a different paradigm.

—Joseph E. Cavanaugh

See also Probability Sampling

Further Reading

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémia Kiadó.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.
- McQuarrie, A. D. R., & Tsai, C.-L. (1998). *Regression and time series model selection*. River Edge, NJ: World Scientific.

Model selection criteria PDF presentations, with an emphasis on AIC and its corrected variants, by Joseph Cavanaugh: http://www.myweb.uiowa.edu/cavaugh/ms_seminar.html

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Weakliem, D. L. (2004). Introduction to the special issue on model selection sociological methods research. *Sociological Methods & Research*, 33(2), 167–187.

The model selection literature has been generally poor at reflecting the deep foundations of the **Akaike information criterion** (AIC) and at making appropriate comparisons to the Bayesian information criterion (BIC). There is a clear philosophy, a sound criterion based in information theory, and a rigorous statistical foundation for AIC. AIC can be justified as Bayesian using a “savvy” prior on models that is a function of sample size and the number of model parameters. Furthermore, BIC can be derived as a non-Bayesian result. Therefore, arguments about using AIC versus BIC for model selection cannot be made from a Bayes versus frequentist perspective. The philosophical context of what is assumed about reality, approximating models, and the intent of model-based inference should determine whether AIC or BIC is used. Various facets of such multimodel inference are presented here, particularly methods of model averaging.

ALCOHOL USE INVENTORY

The Alcohol Use Inventory (AUI; Pearson Assessments, www.pearsonassessments.com/tests/aui.htm) was designed to assess the nature of and problems associated with alcohol use. The AUI is a 228-item self-report inventory that comprises 24 scales, including 17 primary scales, 6 second-order factor skills, and 1 general alcohol involvement scale. These scales provide a basis for describing the multiple dimensions of alcohol use in individuals 16 years of age or older who drink to some extent. It describes the different ways in which individuals use alcohol, such as the benefits they derive from their alcohol use, their style of alcohol use, consequences associated with their use of alcohol, and their degree of concern and acknowledgment of alcohol use.

The AUI is appropriate for individuals who can read at a sixth-grade level or higher. Although the AUI was designed as a self-report inventory, the inventory items can be read to an individual if the respondent cannot read at that level. When taking the AUI, the respondent should be sober and preferably should have been sober for the 8 hours prior to the test. The approximate time for administration is 35–60 minutes. Administration can take longer if the inventory items need to be read to the respondent or if the respondent ponders over items. The AUI can be administered by pencil and paper or by computer. A variety of scoring options are available. The AUI can be scored through computer software, by hand, or through a mail-in scoring service.

Combinations of the previously mentioned scale scores can be used to derive scores and develop typologies and profiles that indicate ways to relate to a client and help with treatment planning. There are some exceptions, however, to the use of the AUI for alcohol-related problems and treatment programs. Even though the AUI was designed to assess the drinking styles of problem drinkers, caution is suggested in using the AUI with individuals convicted of driving under the influence (DUI). It was found that individuals convicted of DUI tended to have lower profiles and that their results should be interpreted with caution.

The AUI was also reported as not appropriate for pre- and posttreatment administration because of the historical nature of the test items. However, the AUI would be appropriate as a baseline measure for a program based on outcomes measurement.

The AUI has a sound basis in research and theory. It was normed on 1,200 individuals who had been admitted to a public in-patient alcohol treatment program. The AUI also demonstrated good reliability and presented evidence of validity. Suggested users of the AUI include psychologists, social workers, chemical dependency counselors, and physicians who work with individuals with alcohol problems. They may find the AUI a useful assessment tool for obtaining information concerning multiple dimensions of problem drinking in their clients.

—Thomas O. Williams, Jr.

See also Reliability Theory; Validity Theory

Further Reading

- Horn, J. L., Wanberg, K. W., & Foster, F. M. (1990). *Guide to the Alcohol Use Inventory (AUI)*. Minneapolis, MN: National Computer Systems.
- Impara, J. C., & Murphy, L. L. (Eds.) (1996). *Buros desk reference: Assessment of substance abuse*. Lincoln, NE: Buros Institute of Mental Measurements.
- Wanberg, K. W., Horn, J. L., & Foster, F. M. (1977). A differential assessment model for alcoholism. *Journal of Studies on Alcohol*, 38, 512–543.

John Leonard Horn: <http://www.usc.edu/projects/nexus/faculty/dept-ldsg/hornjohn/horn.shtml>

ALTERNATE ASSESSMENT

Alternate assessment is a generic term for a family of methods used to assess the academic performance of students with significant disabilities or limited proficiency with English. A small but meaningful number of students have disabilities or limited proficiency with English that make their participation in general state- and district-wide tests impractical, if not impossible, and likely to result in inaccurate measures of

their academic achievement. According to the U.S. Department of Education (USDOE), “An alternate assessment must be aligned with the State’s content standards, must yield results separately in both reading/language arts and mathematics, and must be designed and implemented in a manner that supports use of the results as an indicator of AYP (adequate yearly progress).”

Alternate assessments are an important component of each state’s assessment system and, as such, are required to meet the federal regulations outlined in Title I of the Elementary and Secondary Education Act. Specifically, Title I mandates that “State assessment shall be aligned with the State’s challenging content and student performance standards and provide coherent information about student attainment of such standards” (§1111[b][3][B]). In 2002, the No Child Left Behind (NCLB) legislation increased the federal government’s emphasis on assessment and accountability systems. Specifically, NCLB requires annual statewide assessments for all students in Grades 3–8 and once in high school in reading and language arts, mathematics, and (by 2007) science. Moreover, NCLB requires a disaggregated annual reporting of students’ performance to ensure that all groups (including students with disabilities and English language learners) are making adequate progress toward the goal of all students’ being “proficient” on statewide assessments within the next 12 years.

As noted by Ken Warlick, “The purpose of an alternate assessment should reasonably match, at a minimum, the purpose of the assessment for which it is an alternate. One might ask, ‘If an alternate assessment is based on totally different or alternate standards, or a totally separate curriculum, what is the alternate assessment an alternate to?’”

Alternate Assessments for Students With Disabilities

In 2003, the USDOE reinterpreted the NCLB requirements to allow up to 1% of students in states, school districts, and schools to demonstrate “proficient” performance through participation in statewide alternate

assessment for students with significant cognitive disabilities. However, this interpretation also requires that states’ alternate assessments be reliable and valid measures of students’ achievement of the same rigorous academic content expected of all students. Many states have struggled to meet these requirements because (a) the skills and concepts in the state academic standards were deemed inappropriate or irrelevant for students with significant disabilities, resulting in alternate assessments that focus primarily on functional domains; and (b) the development of the alternate assessment was considered a special education function and therefore only nominally connected to the state’s overall assessment system.

In 2005, the USDOE announced a new policy with respect to students with disabilities as part of the NCLB education reform law. According to this new policy, states may develop modified academic achievement standards and use alternate assessments based on those modified achievement standards for students with disabilities who are served under the Individuals with Disabilities Education Act. States may include proficient scores from such assessments in making AYP decisions, but those scores will be capped at 2% of the total tested population. This provision does not limit how many students may be assessed against modified achievement standards. Individualized education program teams will decide which individual students should take such an assessment.

Like the regulations regarding alternate assessments for students with the most significant cognitive disabilities, the USDOE believes this provision regarding students served under the Individuals with Disabilities Education Act will meet the needs of individual students while ensuring the goals of NCLB are achieved. The provision is intended to allow the success of a student who takes an alternate assessment based on modified achievement standards to be included in calculating school and district performance under AYP. This policy is for those students with disabilities who are expected to make significant academic progress but, because of their disability, are not likely to reach grade-level achievement in the same time frame as all other students.

In order to take advantage of the flexibility provided by the “2% rule,” states are required to develop modified achievement standards and new alternate assessment instruments, provide training and support for individualized education program team members and teachers, and demonstrate that students with disabilities have access to the general curriculum and effective, research-based instruction and intervention. If states meet these requirements, the combination of alternate assessments based on modified achievement standards and alternate assessments for students with significant cognitive disabilities will allow up to 3% of students to demonstrate their proficiency without taking their state’s general large-scale assessment.

Alternate Assessment for English Language Learners

Title I regulations require that English language learners be included in large-scale assessments and accountability “to the extent practicable in the language and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students’ mastery of skills in subjects other than English” [§1111(b)(3); 34C.F.R. 200.4(b)(7)]. Prior to the passage of NCLB, many states and districts exempted from participation in large-scale testing those students who had not been in the United States and in an English language development or bilingual program for at least 3 years. When English language learners are exempted from a general large-scale assessment, however, both NCLB and Title VI mandate that districts and “schools gather information about the academic progress of the exempted students that is comparable to the information from the large-scale assessment.”

Required Characteristics

According to the 2005 USDOE nonregulatory document *Alternate Achievement Standards for Students with the Most Significant Cognitive Disabilities*, alternate assessments must meet standards of high technical quality—validity, reliability, accessibility, objectivity,

and consistency—expected of other educational tests (i.e., *The Standards for Educational and Psychological Testing*, by the American Educational Research Association, 1999). In addition, alternate assessments must have an explicit structure, guidelines for determining which students may participate, clearly defined scoring criteria and procedures, and a report format that communicates student performance in terms of academic achievement standards.

Approaches to Alternate Assessment

Currently, there is no consensus approach to alternate assessment. Three approaches are commonly used and are characterized as (a) portfolios, (b) performance tasks or events, and (c) rating scales. In a review of states’ alternate assessment practices completed by the National Center of Education Outcomes (NCEO), 46% of states indicated they were using some form of portfolio assessments. Performance and portfolio assessments are appealing because of their potential to provide rich descriptions of students’ real-life knowledge and skills. Researchers Browder, Fallin, Davis, and Karvonen, however, have expressed concerns with performance-based approaches and have suggested that the technical characteristics of these alternate assessments may negatively influence students’ and schools’ outcome scores. Initial data from Kentucky’s efforts suggest that reliability of scores may be a source of challenge for states’ portfolio-based alternate assessments. Challenges to the reliability of ratings were also observed by states (e.g., Vermont and Arizona) attempting to use portfolios and performance assessments as part of their general large-scale assessment systems. These difficulties resulted in states’ inability to publicly report assessment results. Moreover, to demonstrate adequate alignment to state standards, performance assessments may need to include numerous tasks and work samples, resulting in an extensive and time-consuming assessment process. Browder and colleagues’ review also identifies student risk factors (e.g., instability of student behavior or health status) as potential influences on students’ alternate assessment results. In the case of on-demand

performance tasks, fluctuations in student behavior or physical well-being could potentially result in inaccurate and invalid assessment results.

Extended Reading and Math Tasks, developed by Tindal and colleagues in 2003, represents a performance task or event approach. Based on curriculum-based measurement technology, this approach consists of a continuum of tasks that measure students' basic skills in reading and mathematics. An extensive literature on the validity and utility of curriculum-based measurement for monitoring students' academic progress provides support for this approach. By including assessment tasks at a range of skill levels, this alternate assessment strategy allows test users to individualize the assessment by administering only those tasks that are considered appropriate to the student's current skills and instructional experiences.

The most recent review of state alternate assessment practices indicates that 30% of states are using a teacher-completed rating scale for their alternate assessment for students with disabilities. A substantial body of evidence on the validity of teachers' judgments of student behavior and academic performance provides support for this approach. In addition, alternate assessments in states using rating scales (e.g., Idaho and Wisconsin) have been judged as adequately aligned to state content standards using the nationally recognized Webb approach to alignment analysis.

Underlying all the approaches to alternate assessment are (a) the collection of classroom-based evidence as indicators of knowledge and skills representative of academic content standards, (b) a scoring rubric for evaluating the knowledge and skills, and (c) a framework for summarizing the level of proficiency exhibited by the collected evidence.

The collection of evidence is a function that consistently is a teacher responsibility. In some cases, evidence is organized in a structured portfolio system that is submitted to a third party for scoring, whereas in others it is loosely organized and remains in the possession of the student's teacher, who is directly involved in scoring. The scoring of the knowledge and skills reflected in the evidence always involves at least two raters, who use an objective rubric to yield

item-level and total test scores. The persons involved in the ratings vary across states; in some cases, educators within the student's school do the scoring; in some cases, scoring is completed in a centralized scoring center; and in still other cases, a combination of local and centralized scoring of evidence is involved. In all cases, significant attention is given to the interrater reliability of scores. In cases in which all the scoring is done within a student's school, states have implemented both preassessment scorer training sessions and postassessment monitoring of evidence collection and scoring practices.

The federal requirements for using the results of an alternate assessment for determining AYP resulted in the need to set proficiency standards for these assessments. Thus, states have had to conduct *standard settings* to generate cut scores for reading and mathematics ratings that correspond to a 3- or 4-level proficiency framework (e.g., minimal performance, basic performance, proficient performance, and advanced performance) commonly used for traditional achievement tests. Perhaps the most important outcome of a standard setting is not the cut scores associated with proficiency levels in each content area but the descriptors of what students who achieve the various performance levels typically know and are able to do. By examining the description of typical student performance in a given performance level, one can gain an understanding of the knowledge, skills, and abilities typically held by students in that performance level and identify things that a given student is not yet able to perform consistently. This type of information helps teachers communicate with others about a student's progress, next year's instructional goals for the student, and the status of the student relative to the state's learning standards.

One area of difficulty is the validity and utility of currently available educational assessments, including alternate assessments. For example, serious questions have been raised about using the results of statewide assessments for (a) monitoring educational performance at the levels of student, classroom, school, and system and (b) making decisions about curriculum and instruction. In the case of English language learners or

students with disabilities, narrowing the enacted curriculum and de-emphasizing other important educational outcomes (e.g., self-determination, social skills, or conversational English) may be unintended consequences of the “new accountability.” Additional research needs to be conducted to determine the curricular validity, instructional utility, and effects of participation in alternate assessments for students with disabilities and for English language learners.

Concluding Points

The development and use of alternate assessments are evolving in a variety of ways across the country. Recent surveys indicate that states are aligning their alternate assessment to varying degrees with the general education curriculum and content standards. The surveys also indicated that a variety of assessment approaches (i.e., direct observation, personal interview, behavioral rating scales, analysis and review of progress, and student portfolios) are being used to evaluate students with severe disabilities. As indicated by the NCEO survey and the authors’ experiences, it appears that a majority of states are borrowing heavily from technology used in the development of behavior rating scales, portfolios, or performance assessments. These technologies are based on teacher observations and the collection of student work samples. These methods, if used appropriately, have the potential to offer statistically sound results. Although relatively little research has been published under the name of alternate assessment, one should not conclude that there is not a research base for alternate assessments. In fact, the conceptual and measurement foundations for alternate assessment are well developed and are based on years of research, in both education and psychology, covering performance assessment, behavioral assessment, developmental assessment, structured observations, and clinical assessment. Although these assessment methods differ somewhat, they all (a) are based on some direct or indirect observation of students, (b) are criterion referenced or domain referenced in nature, and (c) require summary judgments about the synthesis of data and the meaning of the scores or results.

This last quality, the use of judgments by knowledgeable assessors, is the empirical foundation for alternate assessment in many states. A sound research literature exists that supports the fact that teachers can be highly reliable judges of students’ academic functioning.

In summary, information collected through alternate assessments is likely to be different from that collected for students who take large-scale standardized tests, but if it is well aligned with the same academic standards, performance on an alternate assessment can serve as a meaningful index of student progress toward achieving the essential skills and knowledge expected of all students.

—Stephen N. Elliott and Andrew T. Roach

See also Ability Tests; Text Analysis

Further Reading

- Baker, E. L., & Linn, R. L. (2002). *Validity issues for accountability systems*. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities*, 38, 255–270.
- Elementary and Secondary Education Act, 20 U.S.C. 6311(b)(3)(C)(ii) (2002).
- Elliott, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review*, 26, 224–233.
- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105–17, 111 Stat. 37 (codified as amended at 20 U.S.C. § 1400 *et seq.*).
- Korpiva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Roach, A. T., & Elliott, S. N. (2004, April). *Alignment analysis and standard setting procedures for alternate assessments*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., et al. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children*, 69, 481–494.

U.S. Department of Education. (2005, August). *Alternate achievement standards for students with the most significant cognitive disabilities: Nonregulatory guidance*. Washington, DC: Author. Retrieved from <http://www.ed.gov/policy/elsec/guid/altguidance.pdf>

ALTERNATIVE HYPOTHESIS

The term *alternative hypothesis* describes a critical element of hypothesis testing, a popular statistical procedure used by researchers in a wide array of disciplines to evaluate null hypotheses. Although hypothesis testing involves other important elements (such as the level of significance and power), the alternative hypothesis is needed so that the probability associated with the sample data can be computed. If this probability is quite low, the null hypothesis—which initially is presumed to be true—is rejected. Without an alternative hypothesis, there would be no way to compute the sample's probability of occurring, and thus the hypothesis testing procedure would not work.

The Alternative and Null Hypotheses: Similarities and Differences

Both the alternative hypothesis (symbolized as H_1 or as H_a) and the null hypothesis (symbolized as H_0) are statements as to the possible state of affairs in the population(s) of interest. These statements are similar in two other respects: In any given study, both the null and alternative hypotheses must deal with the same statistical concept. Thus, if the null hypothesis deals with the difference between two population means (μ_1 and μ_2), then the alternative hypothesis must also deal with the difference between μ_1 and μ_2 . Moreover, in the usual applied situation, neither H_1 nor H_0 can be proven true on the basis of the study's data.

Although the alternative hypothesis and the null hypothesis are alike in certain ways, they differ in three important ways. First, H_1 and H_0 are “opposites”

in the sense that they say different things about a study's population(s). Second, the hypothesis testing procedure is focused more on the null hypothesis than on the alternative hypothesis. The null hypothesis is always stated first, and it is H_0 that will or will not be rejected after the sample data are analyzed. Finally, it is the alternative hypothesis (and not H_0) that causes a statistical test to be either one-tailed or two-tailed.

Directional and Nondirectional Alternative Hypotheses

The directionality of the alternative hypothesis determines whether a statistical test is conducted in a one-tailed or a two-tailed manner. The alternative hypothesis is said to be directional if it stipulates that the population parameter is positioned on one particular side of the number specified in H_0 . For example, the alternative hypothesis would be directional if it said that a population mean is greater than 20 while the null hypothesis said that 20 is the value of the population mean. Stated symbolically, this situation could be summarized as follows:

$$H_0: \mu = 20$$

$$H_1: \mu > 20.$$

Of course, the alternative hypothesis in this example would also be directional if it were set up to say $H_1: \mu < 20$. Regardless of which way the directional H_1 points, such alternative hypotheses lead to one-tailed tests. This is because the critical region is positioned entirely in one tail of the test statistic's sampling distribution.

The alternative hypothesis is said to be nondirectional if it stipulates that the population parameter is positioned on either side of the number specified in H_0 . For example, the alternative hypothesis would be nondirectional if it says that a population correlation, ρ , has either a positive or negative value while the null hypothesis says that ρ is equal to zero. Stated symbolically, this situation could be summarized as follows:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Nondirectional alternative hypotheses lead to two-tailed tests. This is because the critical region is positioned in both tails of the test statistic's sampling distribution.

It should be noted that two-tailed tests require a larger difference between sample statistics and the hypothesized population parameter in order for the null hypothesis to be rejected. This is the case because the threshold for the critical region will be positioned closer to the center of the test statistic's sampling distribution if that critical region is located entirely in one tail. Because the edges of the critical region are located farther away from the middle of the sampling distribution, some people consider nondirectional alternative hypotheses to be more "conservative" than directional ones.

When H_1 Should Be Specified

For the hypothesis testing procedure to operate properly, the alternative hypothesis must be specified prior to the time any sample data are collected and examined. Unfortunately, some applied researchers violate this rule by switching to a one-tailed test after finding out that they are unable to reject H_0 with H_1 set up in a nondirectional fashion. Such a switch is considered to be a breach of statistical ethics, for the computed probability that is used to decide whether or not H_0 should be rejected is not accurate if H_1 is changed in midstream. (It would be just as "illegal," of course, for a researcher to change H_0 so as to get a desirable result—usually a reject decision—after initially analyzing the data and finding out that a fail-to-reject decision is in the offing.)

—Dong-Ho Park

See also Null Hypothesis Significance Testing; Type I Error

Further Reading

Gold, M. S., Byars, J. A., & Frost-Pineda, K. (2004). Occupational exposure and addictions for physicians: Case

studies and theoretical implications. *Psychiatric Clinics of North America*, 27(4), 745–753.

Shanteau, J. (2001). *What does it mean when experts disagree?* Mahwah, NJ: Erlbaum.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Romero, D. W. (2003). Divisive primaries and the House district vote: A pooled analysis. *American Politics Research*, 31(2), 178–190.

The **alternative hypothesis** in a research study is the one that presents the "foil" or the "educated guess" the researcher makes regarding the exact nature of an absence of a relationship between variables. In this example, the authors point out how the political concept of the divisive primary has fluctuated as investigators have pursued a variety of theoretical and methodological debates. Although most recent studies find that divisive primaries harm general election outcomes, some claim that this effect is spurious, an artifact of uncontrolled electoral prospects phenomena. David Romero argues that this claim is debatable because it rests on questionable conceptual and model constructs and evidence inconsistent with an investigation that controls for the phenomena central to the spurious effect claim. He shows that null and alternative hypothesis findings turn on an unfeatured design characteristic, pooling election years.

AMERICAN DOCTORAL DISSERTATIONS

American Doctoral Dissertations (ADD) is an annual hardcover publication by University Microfilms International (UMI; Ann Arbor, Michigan) for the Association of Research Libraries. As a reference tool, the ADD provides citations to nearly all dissertations written in a given academic year within the United States. In addition to an author index, it provides full bibliographic citations for each dissertation, grouped by subject and by institution. Citations include title of the dissertation, name of the author, degree awarded, awarding institution, year of completion, and UMI order number.

Systematic listing of doctoral dissertations in the United States was begun by the Library of Congress in 1912 in an annual publication titled *List of American Doctoral Dissertations Printed*, which was discontinued in 1938. *Doctoral Dissertations Accepted by American Universities* (1933/34–1955/56) and the *Index to American Doctoral Dissertations* (1955/56–1963/64) were two subsequent annual publications generated by the H. W. Wilson Company. The ADD was introduced in 1934 by Xerox University Microfilms.

The ADD is compiled from the ProQuest Information and Learning database as well as information obtained directly from American universities. The ADD differs from *Dissertation Abstracts International* (DAI) in that it does not include abstracts for each dissertation and does not cover international dissertations. However, the ADD is more comprehensive than the DAI because the ADD includes titles of unpublished dissertations from the ProQuest database.

Currently, the most enhanced version of the ADD and the DAI is ProQuest Digital Dissertations, which is a database of more than 2 million entries for doctoral dissertations and master's theses submitted from more than 1,000 universities and covering the years 1861 to date. ProQuest Digital Dissertations provides the citation, abstract, and first 24 pages of dissertations submitted to UMI within the preceding 2 years. Full versions of these manuscripts and older dissertations may be obtained from UMI in a variety of formats for a fee. Many academic institutions in the United States provide free access to this database for their faculty, staff, and students.

—Marjan Ghahramanlou-Holloway

See also American Psychological Association; Association for Psychological Science

Further Reading

Glatthorn, A. A. (1998). *Writing the winning dissertation: A step-by-step guide*. Thousand Oaks, CA: Corwin Press.

UMI Dissertation Services. (1997). *Publishing your dissertation: How to prepare your manuscript for publication*. Ann Arbor, MI: UMI.

Kirschenbaum, M. G. Electronic theses and dissertations in the humanities: A directory of online references and resources: <http://etext.virginia.edu/ETD/>

UMI ProQuest Digital Dissertations: <http://wwwlib.umi.com/dissertations/>

AMERICAN PSYCHOLOGICAL ASSOCIATION

The American Psychological Association (APA) is the world's largest association of psychologists and is the largest scientific and professional association that represents psychology in the United States. The membership of the APA includes more than 150,000 researchers, educators, clinicians, consultants, and students. The APA's organizational structure includes 55 divisions representing various subfields in psychology and 60 state, provincial, and territorial psychological associations. The association maintains its headquarters in Washington, DC.

Mission

The mission of the APA is to advance psychology as a science and profession and as a means of promoting health, education, and human welfare through the application of science to practice and policy. To achieve this goal, the APA (a) promotes research in psychology and the improvement of research methods and conditions, (b) encourages and advocates for psychology in all its branches and forums, (c) establishes the highest standards of professional conduct and ethics for members of the APA, (d) promotes ongoing improvement of the qualifications and usefulness of psychologists through education and recognition of achievement, and (e) promotes the dissemination of scientific knowledge through meetings, professional contacts, reports, papers, discussions, and publications.

Organizational Structure

The APA governance structure employs a complex system of checks and balances that can be difficult

to comprehend. The APA is chartered in the District of Columbia, and because the charter trumps the association's bylaws, the charter limits what the organization can do in the public policy and advocacy realm to promoting psychology in the public interest. A constitution and bylaws were ratified by the membership more than 50 years ago and remain virtually unchanged today. The primary structural components of the APA include the council of representatives, the board of directors, officers, standing boards and committees, and the central office staff, including a chief executive officer. The members of the APA exercise their power through direct vote and through the election of members to serve on the council of representatives. The primary constituencies from which the representatives are elected are the divisions, which are an integral part of the association, and the state and provincial psychological associations, which are affiliates. The APA divisions include a Division of Evaluation, Measurement and Statistics (Division 5). Much of the work of the Association is done on a volunteer basis by the members of the boards, committees, and ad hoc task forces and working groups. The committees carry out a wide variety of tasks, as indicated by some of their titles: ethics, psychological tests and assessments, membership, and accreditation, to name a few.

The chief executive officer is responsible for the management and staffing of the central office and for running the business aspects of the APA. With nearly 500 employees, the central office provides staff support for all boards and committees; runs one of the largest scientific publishing houses in the world; invests in stocks; manages real estate; and interacts with private, state, and federal agencies and organizations. Member dues represent only 16% of the revenues needed to run the APA.

—Thomas Kubiszyn

See also Association for Psychological Science

Further Reading

APA Web site: www.apa.org

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Finch, S., Thomason, N., and Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, 12(6), 825–853.

The publication guidelines of the **American Psychological Association** (APA) have been the discipline's de facto standards since 1929, and this article documents their advice for authors about statistical practice. Although the advice has been extended with each revision of the guidelines, it has largely focused on null hypothesis significance testing (NHST) to the exclusion of other statistical methods. In parallel, Sue Finch and her colleagues review more than 40 years of critiques of NHST in psychology. Until now, the critiques have had little impact on the APA guidelines. The guidelines are influential in broadly shaping statistical practice although in some cases, recommended reporting practices are not closely followed. The guidelines have an important role to play in reform of statistical practice in psychology. Following the report of the APA's Task Force on Statistical Inference, we propose that future revisions of the guidelines reflect a broader philosophy of analysis and inference, provide detailed statistical requirements for reporting research, and directly address concerns about NHST. In addition, the APA needs to develop ways to ensure that its editors succeed in their leadership role in achieving essential reform.

AMERICAN PSYCHOLOGICAL SOCIETY

See ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

AMERICAN STATISTICAL ASSOCIATION

The American Statistical Association (ASA) is a non-profit organization devoted to the promotion of statistical practice, applications, and research. Its mission includes improving statistical education, fostering excellence in the statistics profession, and enhancing

human welfare. Established in 1839, the ASA currently has 19,000 members in the United States, Canada, and worldwide, and more than 75 chapters and 20 specialized sections. ASA members in government, academia, and the private sector work in diverse areas, including environmental risk assessment, medicine, computing, and social programs.

Formation

The ASA was founded on a commitment to statistical science in service to the public interest, particularly in areas related to public health. The inaugural meeting was held in Boston, Massachusetts, with five founding members: William Cogswell (teacher and genealogist), Richard Fletcher (lawyer and U.S. congressman), John Dix Fisher (physician), Oliver Peabody (lawyer, clergyman, and editor), and Lemuel Shattuck (statistician, genealogist, and publisher). By 1841, the ASA had more than 100 members, primarily in the Boston area; its early membership included Andrew Carnegie, Alexander Graham Bell, and Florence Nightingale. By 1898, the ASA was recognized as a national organization, and membership swelled to more than 500. It is now the largest professional statistical association in the world.

Publications

The ASA publishes refereed journals, books, and newsletters devoted to issues relevant to statistical research and practice. The *Journal of the American Statistical Association* (founded in 1888 as *Publications of the American Statistical Association*) is one of the leading journals in the statistical sciences. *Biometrics Bulletin* was introduced in 1945 to promote the use of statistics in the biological sciences. *Technometrics*, with a focus on statistics applications in the physical, chemical, and engineering sciences, was launched in 1959. In 1976, with the American Educational Research Association, the ASA launched the *Journal of Educational & Behavioral Statistics*. Other ASA publications include the *Journal of Business & Economic Statistics* (1983), the *Journal of Computational & Graphical Statistics* (1992), the

Journal of Agricultural, Biological & Environmental Statistics (1996), and the *Journal of Statistical Education* (1999). An annual *Current Index for Statistics* was introduced in 1975.

Annual Meetings

The ASA hosts annual meetings, symposia, and research conferences. Joint Statistical Meetings are held in conjunction with International Biometric Society, the Institute of Mathematical Statistics, and the Statistical Society of Canada and attract more than 5,000 delegates. Activities of the Joint Statistical Meetings include oral presentations, panel sessions, poster presentations, career placement services, committee meetings, and networking opportunities.

Awards and Educational Programs

The ASA offers research grant programs (cosponsored by the National Science Foundation) and numerous scholarships and awards, including the Statistics in Chemistry Award, the Outstanding Statistical Application Award, and the Gertrude Cox Scholarship, awarded annually to encourage women to enter the statistics professions. Through the Center for Statistics Education, the ASA offers workshops, short courses, and internships, including support for graduate and professional education for teachers of kindergarten through Grade 12.

—Lisa M. Given

See also American Psychological Association; Association for Psychological Science; *Journal of the American Statistical Association*

Further Reading

American Statistical Association Web site: www.amstat.org

AMERICANS WITH DISABILITIES ACT

Patterned largely after Section 504 of the Rehabilitation Act, the Americans with Disabilities Act

(ADA; 42 U.S.C. §§ 12101 et seq.) protects the disabled by imposing far-reaching obligations on private-sector employers, public services and accommodations, and transportation. The ADA provides a comprehensive federal mandate to eliminate discrimination against people with disabilities and provides “clear, strong, consistent and enforceable standards” (§ 12101(b)(2)) for doing so. The ADA’s broad definition of a disability is comparable to the one in Section 504: “(a) a physical or mental impairment that substantially limits one or more of the major life activities; (b) a record of such an impairment; or (c) being regarded as having such an impairment (§ 12102(2)). Further, like Section 504, “[M]ajor life activities” include caring for oneself, hearing, walking, speaking, seeing, breathing, and learning. As with Section 504, the ADA does not require one to have a certificate from a doctor or a psychologist in order to be covered.

The ADA specifically excludes a variety of individuals, most notably those who use illegal drugs (§ 12210). The ADA also specifically excludes transvestites (§ 12208); homosexuals and bisexuals (§ 12211(a)); transsexuals, pedophiles, exhibitionists, voyeurs, and those with sexual behavior disorders (§ 12211(b)); and those with conditions such as psychoactive substance use disorders stemming from current illegal use of drugs (§ 12211(c)). However, the ADA amends Section 504 in that individuals who have successfully completed drug treatment or have otherwise been rehabilitated and are no longer engaged in illegal drug use and who have been “erroneously” regarded as being drug users are covered if they are no longer using illegal drugs (§ 12110). The ADA permits drug testing by employers to ensure that workers are in compliance with the Drug-Free Workplace Act of 1988 (41 U.S.C. Sec. 701). Although it permits employers to prohibit the use of illegal drugs or alcohol in the workplace, the ADA is less clear about the status of alcoholics; it appears that the protections afforded rehabilitated drug users extend to recovering alcoholics.

The ADA addresses three major areas: First, it addresses employment in the private sector and is directly applicable to private schools and colleges. Second, it addresses state and local governments both

as employers and as providers of public services, including transportation, and part of the law applies to public educational institutions. Insofar as the reasonable accommodations requirements in these provisions imply academic program accommodations, qualified students with disabilities can participate in educational institutions at all levels. Third, it deals with private sector public accommodations in buildings and transportation services, and so it may apply to schools and colleges that provide public accommodations. Under its miscellaneous provisions, the ADA stipulates that it cannot be construed as applying a lesser standard than that under Section 504 and its regulations.

—Charles J. Russo

Further Reading

Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101 et seq.

Drug-Free Workplace Act of 1988, 41 U.S.C. Sec. 701 et seq.
Miles, A. S., Russo, C. J., & Gordon, W. M. (1991). The reasonable accommodations provisions of the Americans with Disabilities Act. *Education Law Reporter*, 69(1), 1–8.

Osborne, A. G., & Russo, C. J. (2006). *Special education and the law: A guide for practitioners* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Council for Exceptional Children: <http://www.cec.sped.org>

U.S. Department of Education (updates on regulations, articles, and other general information on the Individuals with Disabilities Education Act and special education): <http://www.ed.gov/offices/OSERS/IDEA/>

U.S. Department of Education, (information from the federal Office of Special Education Programs): <http://www.ed.gov/about/offices/list/osers/osep/index.html?src=mr>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Harrison, T. C. (2002). Has the Americans With Disabilities Act made a difference? A policy analysis of quality of life in the post-Americans With Disabilities Act era. *Policy, Politics, & Nursing Practice*, 3(4), 333–347.

A major challenge in any policy program is to evaluate its effectiveness. One such policy, the

Americans with Disabilities Act (ADA), was signed more than 10 years ago. Policymakers hoped to enable persons with disabilities to combat social barriers such as unemployment by preventing discrimination. It has been the most comprehensive piece of legislation for persons with disabilities in the United States, but the effects of the ADA have been debatable. This article evaluates the effect of the ADA on quality of life for persons with disabilities and offers suggestions for health care policy.

ANALYSIS OF COVARIANCE (ANCOVA)

The analysis of covariance (ANCOVA) can be used to test the null hypothesis of the equality of two or more population means. Alternatively, it can be used in the construction of confidence intervals on differences between means. Although the analysis of variance (ANOVA) is also used for these purposes, ANCOVA has two major advantages over ANOVA in randomized group experiments. First, it generally has higher power. Second, it reduces bias associated with chance differences between groups that exist before the experiment is carried out. These advantages are realized because measurements on one or more nuisance variables are incorporated into the analysis in such a way that (a) the ANCOVA error term is usually smaller (often dramatically so) than the corresponding ANOVA error term and (b) the dependent variable means are adjusted to partially account for chance pretreatment differences between the groups. Hence, nuisance variables play a role in both inferential and descriptive aspects of ANCOVA.

A nuisance variable is defined as a variable that is known to be related to the dependent variable but is of no experimental interest. Suppose, for example, that there is interest in comparing two methods of training workers to perform complex repairs on electronic components; the dependent variable (Y) measures repair proficiency. Two groups are formed using random assignment, and reading skill measurements (X) are obtained. Each group is then exposed to one of

two methods of training. It is known that reading skill is related to performance on the dependent variable, but this relationship is not the focus of the study. Rather, the major focus is whether the two training methods have a differential effect. If some of the within-group variation on the dependent variable is related to reading skill, it is of interest to control for this nuisance variable because it contributes to the error (i.e., within-group) variance. Power is increased whenever a source of nuisance variation is removed from the error variance estimate. This can be accomplished using ANCOVA.

Nuisance variables are usually called *covariates* in the context of ANCOVA. Covariates may be variables that measure constructs that differ from the construct measured by the dependent variable, or they may measure the same construct as the dependent variable does (as in the case of a multiple group pretest-posttest design). In either case, they should be measured before the treatments are applied.

Although ANCOVA is used with several types of research design, it (or the equivalent regression model) is generally most successful with randomized experiments and regression-discontinuity quasi-experiments. Although ANCOVA continues to be widely used in the analysis of observational studies, these designs present special problems that are frequently better handled using other approaches. A strong case can be made for analyzing observational studies using propensity score methods instead of or in combination with modified versions of ANCOVA.

Comparison of ANOVA and ANCOVA

The data presented in Table 1 were collected in a randomized groups pretest-posttest experiment that contained three groups ($n_1 = 10$, $n_2 = 12$, $n_3 = 12$). The purpose of the experiment was to evaluate whether there are differential effects of three training conditions (designated I, II, and III in Table 1) applied to children diagnosed with Down syndrome. Pretest and posttest scores were obtained on a measure known as the Doman-Delacato Profile. The pretest measure was used as the covariate (X), and the posttest was used as the dependent variable (Y). Key differences between

Table 1 Comparison of ANOVA and ANCOVA Summary Tables and Descriptive Statistics Example Data

<i>Treatment</i>					
<i>I</i>		<i>II</i>		<i>III</i>	
<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
35	39.5	52	60	35	39.5
32	35	12	12	48	54
15	18	48	56	44	52
46	54	48	50	18	18
38	42.5	13	15	33.5	36.5
6	10.5	39.5	42	23	23
38	38	17	17	29	33
16	17	38	39.5	9	9
29	32	40	42	32	33
32	35	50	60	37	41
		29	33	32	33

ANOVA summary					ANCOVA summary				
Source	SS	df	MS	F	Source	SS	df	MS	F
Among	253.16	2	126.6	.56	Adj. among	4.64	2	2.32	.47
Within	6536.34	29	225.4		Resid. within	137.25	28	4.90	
Total	6789.50	31			Resid. total	141.89	30		

Sample means		Sample adjusted means	
$\bar{Y}_1$	= 32.15	$\bar{Y}_{1 \text{ adj.}}$	= 35.6
$\bar{Y}_2$	= 38.77	$\bar{Y}_{2 \text{ adj.}}$	= 34.8
$\bar{Y}_3$	= 33.82	$\bar{Y}_{3 \text{ adj.}}$	= 34.7

Notes: I, II, and III represent the training conditions tested; *X* = pretest scores on Doman-Delacato Profile (covariate); *Y* = posttest scores on Doman-Delacato Profile (dependent variable); SS = sum of squares; *df* = degree of freedom; MS = mean square; *F* = Fisher's *F* ratio; adj. = adjusted; Resid. = residual.

ANOVA and ANCOVA applied to these data are described below.

One-Factor ANOVA

Essential descriptive statistics relevant to ANOVA include the sample means on the dependent variable. These means are $\bar{Y}_1 = 32.15$, $\bar{Y}_2 = 38.77$, and $\bar{Y}_3 = 33.82$. An informal inspection of these means (without the aid of inferential methods) might lead one to conclude that the second treatment is the most effective. After all, the mean for this group is almost seven points

higher than is the mean for the first group. But, if one were to also inspect the means for the three groups on the pretest variable *X*, the picture would not seem so clear. These means are $\bar{X}_1 = 28.70$, $\bar{X}_2 = 35.14$, $\bar{X}_3 = 30.95$. Notice that the rank order of the three means on *X* is the same as the order of the three means on *Y*. Further, the sizes of the mean differences are about the same on both *X* and *Y*. The interpretation of the descriptive results on the dependent variable is clouded by the annoyingly large differences between the means on the pretest variable. The pattern of the pretest differences strongly suggests that the posttest

results are simply a reflection of differences that existed before the treatments were carried out. In this example, random assignment resulted in groups that have substantial pretest differences. Because the differences do not appear to be trivial, it is likely that the researcher would like to know the answer to the following question: “What would the means on the posttest have been if the pretest means had been exactly equal?” ANCOVA provides an answer to this question.

One-Factor ANCOVA

Just as the ANOVA F test applies to the means associated with the different treatment conditions, the ANCOVA F test applies to the “adjusted” treatment means. The adjusted means are estimates of what the means on Y would be if all the group means on X were equal to the grand covariate mean. The grand covariate mean is simply the average X score obtained by summing all X scores in the whole experiment and dividing by the total number of X scores. (It is sometimes denoted as $\bar{X} \dots$.) In the case of the example data, $\bar{X} \dots = 31.69$. Hence, ANCOVA attempts to answer the following question: “What would the means on Y have been if each group had a mean score on X of 31.69?” ANCOVA provides the following adjusted mean estimates:

$$\bar{Y}_{1adj} = 35.56, \bar{Y}_{2adj} = 34.84, \text{ and } \bar{Y}_{3adj} = 34.65.$$

A comparison of the unadjusted means with the adjusted means indicates that the adjustment process has substantially changed the means. The adjusted means remove most of the descriptive ambiguity caused by the pretest differences.

Adjustment to the means in this example is substantial, but this is not always the case. The amount of adjustment depends on the size of the mean differences on the covariate and the degree of relationship between the covariate and the dependent variable. There will be no adjustment whatsoever if either (a) there are no differences among the covariate means or (b) there is no linear relationship between the covariate and the dependent variable. Large differences

among covariate means are likely in randomized experiments only if the sample sizes are small (as in the example). Consequently, there is likely to be very little (if any) adjustment of means in large clinical trials or other large randomized groups experiments. But this does not mean that there is no reason to use ANCOVA in place of ANOVA with large experiments. Indeed, the major justification for using ANCOVA rather than ANOVA in randomized experiments is not mean adjustment.

The major reason to prefer ANCOVA over ANOVA is that it is likely to provide a smaller error term. This is important with respect to both the power of hypothesis tests and the width of confidence intervals. These advantages will be large when the within-group linear relationship between X and Y is substantial.

Consider the example data. The error mean square associated with ANOVA on Y is 225, whereas the error mean square associated with ANCOVA is only 4.9. If the power for detecting the maximum difference between the three population means is set at 5 points, the power estimate for ANOVA is only .09; the corresponding power estimate for ANCOVA is more than .99. These data can also be analyzed using other methods of analysis (including the split-plot ANOVA and the one-factor ANOVA applied to change scores), but ANCOVA is usually the method of choice because it generally has higher power. The power advantage of ANCOVA in this experiment is rather dramatic because the pretest (covariate) is very highly correlated with the posttest (dependent variable). Pretests are almost always excellent covariates. There is no requirement, however, that the covariate be a pretest measure. In most randomized groups designs, some easily measured variable that is correlated with but different from the dependent variable is used as the covariate. Sometimes scores on multiple nuisance variables are available. In this case, all nuisance variables can be employed simultaneously as covariates in what is known as a multiple analysis of covariance.

Assumptions and Design Issues

Several assumptions and design aspects are at the foundation of ANCOVA. Strong inferences from

ANCOVA are most easily justified when (a) random assignment is used to form the comparison groups, (b) the covariate(s) is measured before treatments are applied, (c) the individual regression slopes within the groups are homogeneous, (d) the relationship between the covariate and the dependent variable is linear, and (e) the conventional assumptions associated with parametric tests (i.e., independence of errors, homogeneous error distributions, and normally distributed error distributions) are approximately met.

The first two items in this list are easily confirmed design issues. The third issue (homogeneity of regression slopes assumption) should be evaluated whenever ANCOVA is applied. The homogeneity of slopes assumption states that the slope of Y on X is the same within each individual treatment population. When the individual slopes are not the same, both the descriptive and the inferential aspects of the analysis are suspect. If the slopes are not the same, the size of the treatment effect is a function of the value of the covariate, but the results of ANCOVA will not acknowledge this important fact.

When any of the design aspects or assumptions are incompatible with the nature of the study, remedial solutions are available. These include modifications of ANCOVA and alternative methods of analysis that either correct for the problems or are less sensitive to them.

—Bradley E. Huitema

See also Analysis of Variance (ANOVA)

Further Reading

- Huitema, B. E. (in preparation). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and observational studies* (2nd ed.). Hoboken, NJ: Wiley.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- McKean, J. W., & Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *American Statistician*, 48, 220–229.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.

Visual Statistics with Multimedia, an online tutorial that covers ANCOVA: <http://pages.infinit.net/rlevesqu/spss.htm>

Web-based software that performs both traditional ANCOVA and a more recently developed, robust ANCOVA (based on the work of McKean & Vidmar, 1994): www.stat.wmich.edu/slab/RGLM/. Click the Online Resources button under the Site Guide, and then click on RGLM.

ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance (ANOVA) was developed by Ronald A. Fisher in the 1930s (although the name “analysis of variance” came later from John W. Tukey). ANOVA refers to a family of statistical procedures that use the F test to test the overall fit of a linear model to the observed data. Although typically associated with the analysis of experimental research designs in which categorical independent variables are manipulated to see the effect (if any) on a continuous dependent variable, these designs are merely special cases of a general linear model in which the categorical independent variables are expressed as dummy variables. As such, ANOVA embodies a family of tests that are special cases of linear regression in which the linear model is defined in terms of group means. The resulting F test is, therefore, an overall test of whether group means differ across levels of the categorical independent variable or variables.

Different Types of ANOVA

ANOVA can be applied to a variety of research designs and takes specific names that reflect the design to which it has been applied. The computational details of the analysis become more complex with the design, but the essence of the test remains the same. The first distinction that is made is in the number of independent variables in the research design. If there is simply one independent variable, then the ANOVA is called a one-way ANOVA. If two independent variables have been manipulated in the research, then a two-way ANOVA can be used to analyze the data; likewise if three independent variables

have been manipulated, a three-way ANOVA is appropriate. The logic of the test extends to any number of independent variables; however, for ease of interpretation, researchers rarely go beyond a three-way ANOVA.

The second distinction that needs to be made is whether data in different conditions are independent or related. If data representing different levels of an independent variable are independent (i.e., collected from different entities), then an independent ANOVA can be used (also known as a between-groups ANOVA). If two independent variables have been used and all levels of all variables contain data from different entities, then a two-way independent ANOVA could be employed, and so on. When data are related—for example, when different entities have provided data for all levels of an independent variable or all levels of several independent variables—then a repeated measures ANOVA (also known as within-subjects ANOVA) can be employed. As with independent designs, it is possible to have one-way, two-way, three-way, n -way repeated measures ANOVAs. A final type of ANOVA is used when a mixture of independent and related data have been collected. These mixed designs require at least two independent variables, one of which has been manipulated using different entities (and so data are independent) and the other of which has been manipulated using the same entities (data are related). In these situations, a mixed ANOVA is used. It is possible to combine different numbers of independent variables measured using different entities or the same entities to come up with three-way, four-way, or n -way mixed ANOVAs. ANOVAs involving more than one independent variable are known as factorial ANOVAs.

Similarities Among Different ANOVAs

All the ANOVAs above have some common features. All of them produce F tests that are the ratio of the variance explained or accounted for by a particular effect compared to the variance that cannot be explained by that effect (i.e., error variance). The computational details of a simple ANOVA are described in the entry titled “One-Way Analysis of

Variance.” In experimental scenarios, the F test can be thought of as the ratio of the experimental effect to the individual differences in performance. The observed value of F is compared with critical values of F from a special distribution known as the F distribution, which represents the values of F that can be expected at certain levels of probability. If the observed value exceeds the critical value for a small probability (typically 0.05), we tend to infer that the model is a significant fit of the observed data or, in the case of experiments, that the experimental manipulation has had a significant effect on performance.

Differences Among ANOVAs

The main difference among ANOVAs is the effects that they produce. In an ANOVA with one independent variable, a single value of F is produced that tests the effect of that variable. In factorial ANOVAs, multiple F s are produced: one for each effect and one for every combination of effects. The entry “One-Way Analysis of Variance” describes an example about the effect of mood induction on the number of items people would generate when asked to list as many items as they could that needed checking before the people left on holiday. This experiment involved groups reflecting different levels of the independent variable: negative mood, positive mood, and no mood induction. In the study proper, a second independent variable related to whether participants were instructed to generate as many items as they could or to generate items until they felt like stopping. This second independent variable could be called the *stop rule* and had two levels: “as many as can” and “feel like stopping.” This experiment requires a factorial ANOVA, and the result would be an F ratio for the effect of mood (this is known as a *main effect*), a different F ratio for the main effect of the stop rule, and a third F ratio representing the combined effect of mood and the stop rule, known as the mood-by-stop-rule interaction.

Regardless of whether the factorial ANOVA is independent, repeated measures, or mixed design, the result is the same: F associated with each main effect, and F s associated with each interaction term. Sticking with the

example above, if we added a third variable, such as gender, into the design, we would end up with three main effects: (a) mood, (b) stop rule, and (c) gender.

Three interaction terms involving two variables (known as two-way interactions) would result in the following main effects:

1. Mood $\times$ Stop rule
2. Mood $\times$ Gender
3. Stop rule $\times$ Gender

One interaction of all three variables (known as a three-way interaction) would result in one main effect:

$$\text{Mood} \times \text{Stop rule} \times \text{Gender}$$

Each of these effects would have an associated *F* ratio that tested whether the effect had an influence on the group means. The derivation of these *F*s is affected by whether the design is repeated measures, independent, or mixed, but the interpretation of these *F*s is unaffected by the design.

Follow-Up Tests

Unless a main effect represents a difference between two groups (such as the main effect of the stop rule, above), the *F* tells us only that the groups' means differ in some way (across one or more variables, depending on whether it is a main effect or an interaction). Main effects are usually followed up either with planned comparisons, which compare specific sets of means, or with post hoc tests, which compare all combinations of pairs of means (see, for example, the entries on Bonferroni Test and Newman-Keuls Test). In factorial designs, the interactions are typically more interesting than the main effects. Interactions are usually broken down using simple effects analysis or specific contrasts designed by the researcher.

ANOVA as a General Linear Model

When ANOVA is used to analyze data from groups, it is a special case of a linear model. Specifically, the

linear model can be expressed in terms of dummy variables. Any categorical variable can be expressed as a series of 0s and 1s; there will always be one less variable than there are groups, and each variable compares each group against a base category (e.g., a control group). The example from the entry "One-Way Analysis of Variance," described above (ignoring the second independent variable of the stop rule, to keep things simple), can provide an illustration. Remember that this experiment involved groups reflecting different levels of the independent variable: negative mood, positive mood, and no mood induction. This scenario can be represented by a standard regression equation:

$$\text{Items Generated}_i = b_0 + b_1 \text{Negative Mood}_i + b_2 \text{Positive Mood}_i + \varepsilon_i$$

in which Negative Mood is a binary variable coded 1 for people undergoing a negative mood induction and 0 for all other groups, and Positive Mood is a binary variable coded 1 for the positive mood induction group and 0 for all other groups. The control group (no mood induction) is coded zero for both variables. It turns out that b_0 represents the mean of the control group (i.e., the mean number of items generated when no mood induction is performed); b_1 is the difference between the mean number of items generated when a negative mood induction is done and the mean number of items generated when no mood induction is done; and b_2 is the difference between the mean number of items generated when a positive mood induction is done and the mean number of items generated when no mood induction is done. More complex designs such as factorial ANOVA can be conceptualized in a similar way.

Assumptions

For the *F* ratio to be accurate, the following assumptions must be met: (a) observations should be statistically independent, (b) data should be randomly sampled from the population of interest and measured at an interval level, (c) the outcome variable should be sampled from a normal distribution, and (d) there must be homogeneity of variance.

Differences With Repeated Measures

When data are related (i.e., when the independent variable has been manipulated using the same entities), the basic logic described above still holds true. The resulting F can be interpreted in the same way, although the partitioning of variance differs somewhat. However, when a repeated measures design is used, the assumption of independence is violated, giving rise to an additional assumption of *sphericity*. This assumption requires that the variances of difference scores between conditions be roughly equal. When this assumption is not met, the degrees of freedom associated with the F value must be corrected using one of two estimates of sphericity: the Greenhouse-Geisser estimate or the Huynh-Feldt estimate.

—Andy P. Field

See also Analysis of Covariance (ANCOVA); Bonferroni Test; Dependent Variable; Factorial Design; Fisher, Ronald Aylmer; Homogeneity of Variance; Independent Variable; Linear Regression; Newman-Keuls Test; One-Way Analysis of Variance; Tukey-Kramer Procedure; Variance

Further Reading

- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6), 426–443.
- Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). Perseveration of checking thoughts and mood-as-input hypothesis. *Journal of Behavior Therapy & Experimental Psychiatry*, 34, 141–160.
- Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Castañeda, M. B., Levin, J. R., & Dunham, R. B. (1993). Using planned comparisons in management research: A case for the Bonferroni procedure. *Journal of Management*, 19(3), 707–724.

This article describes the Bonferroni multiple-comparison procedure (used in conjunction with the robust F test) and makes a case for researchers' more frequent and appropriate use of it. The procedure is discussed as a test that facilitates

investigation of precise and powerful a priori multiple comparisons. Characteristics of the Bonferroni procedure are described in relation to the more familiar Scheffe post hoc multiple-comparison method, and a step-by-step guide for comparing and choosing between the two is provided. The Bonferroni procedure is discussed in detail in the context of one-factor **analysis of variance (ANOVA)** designs. Application of the technique is then considered in the context of factorial designs, analyses of covariance, univariate repeated measures analyses, multivariate ANOVAs, and recent sequential hypothesis-testing extensions.

ANTHROPOMETRY

Anthropometry is the measurement of the human body. It is distinct from osteometry, which is the measurement of skeletal material. Anthropometry is sometimes subdivided into craniofacial anthropometry (measurement of the head and face) and somatometry (measurement of the body). Two-dimensional measurement of the head from x-ray cephalograms is known in the United States as cephalometry. In Europe, on the other hand, cephalometry refers to measurement of the head and face, while measurement of x-ray tracings is known as roentgen-cephalometry.

Canons, or simple rules of proportionality based on multiples of specific body parts, were used by classical Greek, Roman, and Renaissance artists to describe the shape of the human body and were based on aesthetic ideals rather than measurement. Anthropometry, which uses actual body measurements, did not develop until 1654, when a German anatomist at the University of Padua, Johann Sigismund Elsholtz, developed a standardized measuring tool for his doctoral dissertation on the symmetry of the human body. He created a vertical rod divided into six equal parts, which he called *pedis* (feet). He then subdivided each foot into twelve equal parts, which he called *uncias* (inches). This “anthropometron” is virtually identical to the modern anthropometer, used in most doctors' offices for measuring height.

After graduation, Elsholtz abandoned anthropometry for research in botany, herbal medicine, distillation, and intravenous infusion. However, his technique was adopted widely in the 18th century in early studies of human growth and development. In the early 19th century, the applications of anthropometry expanded to include measurements used to classify human populations on the basis of quantitative morphology. This research grew out of an interest in Linnaean systematics and taxonomy, with its emphasis on typology and “ideal” types to define contemporary populations.

More sophisticated measuring instruments, including spreading and sliding calipers, were devised to measure the human body, especially the skull, in greater detail than was possible with Elsholtz’s anthropometron. Numerous landmarks on the head and body were identified for measuring an ever-increasing number of linear and contour dimensions. These early techniques were highly idiosyncratic, with researchers using their own measurement system. Ambiguity in the names and descriptions of the landmarks and confusion as to the actual measurements being taken made interobserver error a serious problem when comparing anthropometric measurements taken by different researchers. For example, one of the most basic measurements, maximum cranial length, is also known as head length, maximum glabello-occipital length, maximum head length, *diamètre antero-postérieur maximum ou glabellaire*, or *grösste Kopflänge*. This measurement usually is taken between the landmarks of the glabella, which is defined as the most prominent point in the median sagittal plane between the supraorbital ridges, and the *opisthocranium*, the most prominent posterior point in the median plane of the occiput, or back of the skull. Both landmarks have numerous synonyms. Glabella also is known as the nasal eminence or *bosse moyen*, while *point occipital maximum* and *extremum occiput* are synonyms for opisthocranium.

Much of the early anthropometric research focused on the Cephalic Index, the ratio of cranial width to length, developed by Swedish anatomist Anders Retzius to classify living populations according to head shape. Variations in measurement technique,

plus disparities in classification systems, resulted in a bewildering variety of categories in this index. These differences produced so much confusion about this simple ratio that French anthropologist Paul Topinard devoted an entire chapter of his textbook on anthropology to the Cephalic Index.

By 1870, a consensus had developed around the work of French anatomist Paul Broca. However, the emergence of the modern German state following the Franco-Prussian War of 1870 led to the establishment of a separate, distinctly German school of anthropometry, which was formalized by the Frankfurt Convention of 1882. The Convention had one long-term scientific result, the establishment of the Frankfurt horizontal plane, a standard reference line connecting the upper edge of the ear canal and the inferior rim of the orbit. This plane is still used in anthropometric and cephalometric studies to ensure standardization of head position while measurements are taken.

Following the Convention, several international attempts were made to standardize anthropometry, but these collapsed by the beginning of World War I. Two textbooks developed out of these efforts. The first was Rudolf Martin’s *Lehrbuch der Anthropologie*, the standard reference for the German school. In 1920, Aleš Hrdlička, one of the founders of American anthropology, wrote *Anthropometry*, based on his studies in France, to set the North American standards in the field. Hrdlička also was the first to propose the use of anthropometry in medicine.

As anthropometry developed during the 19th century, anthropometric data often were misapplied by scientists of the day to substantiate racial, class, and gender stereotypes. For example, during the period 1820 to 1851, an American scientist, Samuel George Morton, collected more than 1,000 human skulls to measure their cranial capacity as a way to rank races. (At the time, it was mistakenly assumed that a large cranial capacity equated with increased intelligence.) Morton would scientifically measure the skulls but would bias his sample by omitting individuals or groups that would not prove the superior cranial capacity of the white race. Cesare Lombroso (1835–1909), an Italian physician and criminologist, claimed that he

could identify criminals using anthropometric features that reflected what he considered to be “atavistic” traits (or traits that demonstrate a reappearance of ape-like characteristics). Some of the “criminal characteristics” he used were shape and size deviations in the head; large jaws, teeth, and cheekbones; long arms; and protruding lips. The measurements themselves were not faulty. Instead, the anthropometric traits used in these pseudoscientific studies were carefully chosen to reinforce the prejudices of the day.

Anthropometry as a technique for studying human populations fell out of favor at the beginning of the 20th century. One reason was a reaction against the misuse of anthropometric data in racial studies. Other causes were the discovery of blood group genetics, which provided a much more precise way to study populations, and the recognition that the Linnaean “types” anthropologists studied were not fixed but were subject to environmental influences. This was based in part on Hrdlička’s research on growth in immigrant families, which demonstrated that the children of first-generation immigrants were nearly always taller than their parents. The genetics of the immigrant children had not changed, but their nutritional status had improved. Also, the development of x-ray technology permitted more detailed two-dimensional measurements to be taken of skeletal features than could be obtained through direct measurement.

Although racial typology lost favor as a subject of anthropometric research, the study of human growth and development has continued, with anthropometry the primary means of evaluation. Much of the emphasis has shifted from academic studies to applied research. Following the advice of Hrdlička, anthropometry now is used widely in the medical field. Measurements of height and weight provide data for the calculation of body composition, which assists in the assessment of nutrition, physiology, growth, and development, as well as adding to the understanding of human obesity. Craniofacial measurements are used to analyze the patterns of dysmorphology associated with a wide range of congenital craniofacial anomalies, calculate the quantitative surgical changes needed to improve the deformities, and evaluate post-operative growth and outcomes.

Detailed measurement of the human body and its segments is used in the field of kinanthropometry to analyze the relationship between anatomy and movement, an application of particular importance in sports medicine. Anthropometry is used extensively in human engineering, or ergonomics, to study the fit between human morphology and the physical environment in which humans function. This field ranges from the fit of clothing, especially personal protective equipment, to the design of furniture, living and working spaces, and transportation.

Both somatometry and craniometry are employed in forensic science as tools for human identification. Body measurements are correlated with skeletal dimensions to create regression equations that estimate body height, determine sex, determine age in subadults, and identify possible racial affinities. Craniofacial measurements aid in reconstructing the facial features of skeletal remains and in age enhancement of photographs of missing children.

Psychologists use anthropometry to study the mechanisms of facial recognition, increasingly important in security issues. Other research attempts to quantify facial attractiveness. Some researchers, like Elsholtz some 350 years ago, have focused on the influence of facial symmetry, and others have identified the importance of statistical averages in defining attractiveness, updating the research of Sir Francis Galton on composite photographs.

The increasing use of anthropometry has led to further expansion of measurement instruments. Measurements of the body can be taken directly through spreading and sliding calipers, skin fold calipers, tape measures, and weight scales and can also be obtained from radiographs, dual energy x-ray absorptiometry scans, CAT scans, and other radiographic techniques. With direct measurement of the body, there is an inherent problem with body contours. In contrast to a dry skull that has many observable landmarks, such as sutures, which reflect where the bones unite, the living body is covered with skin and therefore has fewer easily discernable landmarks for measurement. On the body of someone who has anomalies due to disease or illness, the difficulty in locating landmarks is further exacerbated.

The lack of standardization in anthropometry remains a major problem. In even the most basic anthropometric texts, the definition of the landmarks used for measurement, the instruments to use, and the best methodology to take a given measurement often remain unclear. It is difficult to improve on an accuracy of better than 5 mm in most body measurements. In an application such as craniofacial reconstructive surgery, a measurement error this large would jeopardize the surgical outcome, so the cranial landmarks, synonyms, instruments, and techniques are more elaborately described, along with a discussion of variables to consider when taking the measurements, to minimize measurement error.

Care must be taken with measurement techniques. For example, individual height decreases by 1–3 mm during the day, so to track the growth of a child accurately, height should be measured at the same time of day. Height also should be consistently measured with an individual in stocking feet. Adult height decreases with age as a result of compression of intervertebral discs and bone thinning and cracking due to diseases such as osteoporosis. Other factors that affect measurement accuracy are well-calibrated instruments; replacement of instruments such as tape measures, which stretch with repeated use; standardized positioning of the body, such as the Frankfurt horizontal plane in cranial measurements, to eliminate error; a good source of natural light for ease of reading the instruments; a sequence for ease of taking the measurements; and finally, a good recording system to track the measurements.

With thousands of anthropometric measures available, proper scientific study using these measurements depends on the careful selection of the most meaningful measures to eliminate wasted research time collecting irrelevant data. With small samples, the data may have to be standardized before any analysis can begin to ensure that the results are meaningful. In studies based on larger groups of individuals, results can be broken down easily by sex, age, population of origin, health status, and other categories to provide reliable data for subsequent applications.

With improved techniques, anthropometry has been resurrected from its earlier confusion and

misapplication to become an integral component in fields as diverse as forensic identification, growth and development studies, reconstructive surgery, ergonomics, and nutritional evaluation.

—Elizabeth M. Salter and John C. Kolar

See also Measurement; Measurement Error

Further Reading

- Elsholtz, J. S. (1654). *Anthropometria*. Padua: M. Cadorini.
- Galton, F. (1878). Composite portraits. *J Anthropol Inst Gr Brit & Ireland*, 8, 132.
- Garson, J. G. (1884). The Frankfort Craniometric Convention, with critical remarks thereon. *J Anthropol Inst Gr Brit & Ireland*, 14, 64–83.
- Gould, S. J. (1996). *The mismeasure of man* (rev. & exp. ed.). New York: Norton.
- Kolar, J. C., & Salter, E. M. (1997). *Craniofacial anthropometry. Practical measurement of the head and face for clinical, surgical and research use*. Springfield, IL: Charles C Thomas.
- Lombroso, C. (1876). *L'uomo delinquente*. Milan, Italy: Hoepli.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Yin, Z., Hanes, J., Jr., Moore, J. B., Humbles, P., Barbeau, P., & Cutin, B. (2005). An after-school physical activity program for obesity prevention in children. *Evaluation & the Health Professions*, 28(1), 67–89.

This article describes the process of setting up a 3-year, school-based afterschool physical activity intervention in elementary schools. The primary aim of the study is to determine whether adiposity and fitness will improve in children who are exposed to a “fitogenic” versus an “obesogenic” environment. Eighteen schools were randomized to the control (obesogenic) or intervention (fitogenic) group. The intervention consisted of (a) academic enrichment, (b) a healthy snack, and (c) physical activity in a mastery-oriented environment, and outcome measures were **anthropometry**, body composition, blood samples, psychological tests, and other measures. Successful implementation would show the feasibility of schools’ being able to provide a fitogenic environment, and significant differences between the groups would provide evidence that a fitogenic

environment after school has positive health benefits. If feasibility and efficacy are demonstrated, implementing an afterschool program like this one in elementary schools could play a major role in preventing and reducing childhood obesity.

APPLIED RESEARCH

Whereas basic research is the study of fundamental principles and processes, the defining characteristic of applied research is that its research findings have immediate application to the general topic under consideration.

Another characteristic of applied research is that its goal is to solve practical questions, and in contrast to basic research, applied research is aimed not at understanding or accumulating more knowledge about some phenomenon but at describing the phenomenon. While applied research has a more pragmatic element than basic research does, basic research forms the foundation for applied research.

For example, reading is a skill that involves many different processes, including visual and intellectual skills. The basics of how the eyes focus on letters and words and how that message is transmitted to the brain and then translated into meaningful symbols may very well constitute a set of basic research questions. In contrast, an example of an applied research endeavor is taking those findings and using them to understand why some children read better than others or creating an intervention and providing it for a group of children who are poor readers.

—Neil J. Salkind

See also Basic Research

Further Reading

Cochran, E. L. (1994). Basic versus applied research in cognitive science: A view from industry. *Ecological Psychology*, 6, 131–135.

Basic vs. applied research (discussion from the Ethical, Legal, and Social Issues in Science site at the University of California): <http://www.lbl.gov/Education/ELSI/research-main.html>

APTITUDE TESTS

The term *aptitude*, according to most dictionaries, is derived from the Latin term *aptitudo*, meaning fitness. The psychological use of the term is similar in that it has traditionally referred to a potential for acquiring knowledge or skill. Traditionally, aptitudes are described as sets of characteristics that relate to an individual's ability to acquire knowledge or skills in the context of some training or educational program. There are two important aspects of aptitude to keep in mind. First, aptitudes are present conditions (i.e., existing at the time they are measured). Second, there is nothing inherent in the concept of aptitudes that says whether they are inherited or acquired or represent some combination of heredity and environmental influences. Also, aptitude tests do not directly assess an individual's future success; they are meant to assess aspects of the individual that are indicators of future success. That is, these measures are used to provide a probability estimate of an individual's success in a particular training or educational program. While the meaning of *aptitude* is well delineated, there is much controversy over how to distinguish aptitude tests from other kinds of psychometric measures, specifically intelligence and achievement tests, partly because the major salient difference between intelligence, aptitude, and achievement tests has to do with the purpose of testing rather than with the content of the tests. What makes an assessment instrument an aptitude test rather than an intelligence or achievement test is mainly the future orientation of the predictions to be made from the test scores.

Historians generally date the movement of modern psychological testing from the 1905 work by Alfred Binet and Théodore Simon in developing a set of measures to assess intelligence. The Binet-Simon measures, and especially the English translation and refinement made by Lewis Terman in 1916, called the Stanford-Binet, are in widespread use even today. Few adults living in industrialized countries today have avoided taking at least one test of intelligence during their school years. Intelligence tests were designed with the goal of predicting school success.

Thus, in terms of the definition of aptitude provided above, when the purpose of an intelligence test is prediction, then the intelligence test is essentially an aptitude test—although an aptitude test of general academic content (e.g., memory, reasoning, math, and verbal domains). Aptitude tests, however, sample a wider array of talents than those included in most general intelligence measures, especially in the occupational domain. By the late 1910s and early 1920s, dozens of different aptitude tests had been created for prediction of success in a variety of different occupations (e.g., auto mechanic, retail salesmen, waitress, telegrapher, clerk, Hollerith operator, musician, registered nurse).

It is important to distinguish between so-called trade tests and aptitude tests. The distinction rests more on the characteristics of the examinee population than on the content of the tests. That is, when all the examinees can be expected to have similar prior exposure to the knowledge and skills needed to perform well on the test, the test is essentially one of ability or aptitude. But when prior knowledge and skills have an important impact on the examinees' success on the test, it is essentially an achievement test, or a measure of learned knowledge or skills, rather than an assessment of potential for acquiring such knowledge or skills. For psychologists who design aptitude tests, this is a critical concern. For example, the psychologist must be able to determine whether reading skills are an important determinant of test performance in order to present the test material in a paper-and-pencil format. Intelligence test developers assumed that individual differences in reading skills in young children were possible confounding influences, and so the developers created intelligence tests that did not require a child to know how to read or write. For assessing the aptitude of adults for an office clerk job, however, being able to read would be a prerequisite skill, so a paper-and-pencil aptitude test would certainly be appropriate.

Utility of Aptitude Tests

Aptitude tests are useful for the purpose of aiding educational or occupational selection when there are

marked individual differences in the likelihood of success that are, in turn, determined by cognitive, perceptual, or physical abilities. The degree of utility of an aptitude test is determined by three major factors: (a) the cost of training or education, (b) the correlation between the aptitude test scores and success on the educational or occupational criterion, and (c) the ratio of the number of applicants to the number of places to be filled. When training is expensive, the cost to the organization of having trainees fail can be an important factor in adopting an aptitude testing program for screening applicants. When training is brief or inexpensive, such as for retail sales or other entry-level positions, the value of aptitude testing is diminished because the cost of accepting applicants who fail is not as burdensome for the organization. The correlation between aptitude test scores and success measures will determine how accurate the prediction of success or failure is. The larger the correlation, the more accurate the prediction. Finally, when there are many more applicants than spaces to be filled, the aptitude test will be more effective in maximizing the overall success rate. In contrast, when there are few applicants for each position, and thus nearly all applicants are accepted, the ranking of applicants by aptitude becomes largely irrelevant.

Two Types of Aptitude Tests

The aptitude tests developed over the past century have generally bifurcated into two different types: job-specific tests and multiaptitude batteries. Similar to the early aptitude tests described above, job-specific aptitude tests are typically designed to determine which candidates are best suited to particular occupations. In theory, there can be as many different occupational aptitude tests as there are differentiable occupations. In practice, however, there are common aptitudes underlying many occupations. For example, different kinds of mechanical jobs (e.g., auto mechanic, electronics service repair, assembly worker) may all involve aptitudes for dexterity, fine motor coordination, visual perception, and so on. An organization that wishes to select employees for a particular occupational placement might attempt to

identify (through job analysis) what particular aptitudes are needed for successful job performance. The organization, in order to select the applicants who are most likely to succeed in a training program, can then create an aptitude measure that samples these specific aptitudes. Alternatively, among the dozens of commercially available tests, the organization may find an off-the-shelf aptitude measure that covers the most important aptitudes for training success for the particular job.

The other kind of aptitude measure is the multiaptitude battery. These tests are used frequently in educational contexts, and some are used in large-scale employment testing situations. In the educational context, multiaptitude tests may be very general, such as the Scholastic Aptitude Test, which was created in 1926 for selecting high school students for college and university placement. Today, the Scholastic Aptitude Test is one of the most widely used aptitude test batteries in the United States and is administered to more than 1 million students each year. The original Scholastic Aptitude Test assessed only two broad academic aptitudes: verbal and math. The most recent modification of the Scholastic Aptitude Test also includes a writing component. Multiaptitude test batteries can also be designed to provide assessments across several different aptitudes. The first large-scale multiaptitude batteries for use in educational contexts were developed by Thurstone and Thurstone in the early 1940s and became known as the Primary Mental Abilities battery. Another battery, the Differential Aptitude Tests (DAT), was introduced in 1947 and is still in use today. The DAT provides scores on eight different aptitudes (verbal, numerical, abstract reasoning, clerical speed and accuracy, mechanical reasoning, spatial relations, spelling, and language use).

There are many more such multiaptitude batteries that are administered in schools throughout the United States each year. Many of these tests do not have the term *aptitude* in their titles, but they are similar in content coverage and in the general purposes of testing. Such educational aptitude batteries are primarily used for counseling purposes. That is, the underlying premise for the utility of these tests is that they allow a parent or counselor to identify an individual

student's aptitude strengths and weaknesses. Usually, the test information is presented as a profile, a set of bar graphs that show where the student stands in respect to some norming group on each of the different aptitudes. Counselors may use this information to help guide the student in a way that either builds on the student's strengths or attempts to remediate the student's weaknesses. In practice, however, many of the different aptitudes assessed with these measures are themselves substantially positively correlated because of shared variance with general intelligence. When that happens, it is more difficult to provide a reliable differentiation among the individual's strengths and weaknesses. This is one of the most intractable problems associated with the counseling use of multiaptitude test batteries.

Multiaptitude batteries for occupational selection tend to be somewhat more useful for selection and classification purposes. (Classification is the process of assigning particular individuals to specific jobs by matching the individual's profile of aptitude strengths and weaknesses to the job requirements.) The two largest occupational multiaptitude test batteries used in the United States are the Armed Services Vocational Aptitude Battery (ASVAB) and the General Aptitude Test Battery (GATB). The ASVAB is used by the U.S. armed forces, and until recently, the GATB was used by federal and state employment agencies. In contrast to the multiaptitude batteries described above for educational contexts, these two tests are explicitly linked to a wide variety of specific occupations. For example, when individuals complete the ASVAB, they are each provided with a set of scores that determines their suitability for all the different entry-level occupations within the military. With that information, they can be classified into the occupation in which they are most likely to succeed.

Concerns About Aptitude Tests

Although aptitude tests have been shown to be quite effective predictors of future academic and occupational performance, they have been somewhat controversial because of the meaning inherent in the

assessment of *potential* and because of a wide variety of group differences in performance on standardized aptitude tests. Experience with the Scholastic Aptitude Test, for example, has indicated marked mean score differences between male and female test takers; between black, white, Hispanic, and Asian-American test takers; and between socioeconomic status groups. Because the Scholastic Aptitude Test is not traditionally considered to be taken by a representative or random sample of 16–18 year olds (since those students taking the test essentially are self-selected college-bound individuals), group differences on the Scholastic Aptitude Test do not provide direct evidence for overall group differences in academic potential. However, the differences between group means are significant and sometimes substantial, which has led many commentators to question whether and how much the test is associated with prior educational background and other demographic variables. Much of the difficulty centers around the term “potential” associated with aptitude tests, in contrast with achievement measures. That is, if these different groups differ only in terms of academic achievement, there would be perhaps less controversy than there is if the groups are determined to differ in terms of academic potential. Many testing organizations have in fact revised the names of their aptitude tests to remove the term that is associated with potential (e.g., the Scholastic Aptitude Test became the Scholastic Assessment Test in the 1990s). At one level, such a change may be cosmetic, but at another level, it does show that testing organizations have come to recognize that one does not need to imbue a test with the notion of potential in order to make predictions about future academic or occupational performance. That is, there is nothing inherently problematic in using an intelligence or achievement test for the same purpose as an aptitude test as long as it taps the same underlying knowledge and skills that are critical for performance on the predicted criterion measure. Given that intelligence, aptitude, and achievement tests assess only current performance, it is ultimately the prediction aspect of a test that makes it an aptitude test. Furthermore, it is fundamentally impossible to know what an individual’s actual

potential is for academic or occupational knowledge or skills, because it is not possible to know what the universe of instructional or training programs may be. Should methods of instruction or training be improved at some time in the future, even those individuals with relatively lower aptitudes may show marked increases in performance. In that sense, the operational conceptualization of aptitude has to be in terms of whatever instructional or training methods are actually in use at any one time.

Over- and Underachievement

One aspect of aptitude tests that has been very much misunderstood is the notion of over- and underachievement. Typically, the term *overachiever* is given to individuals who have relatively higher scores on achievement tests than they do on aptitude tests, and the term *underachiever* is given to individuals who have relatively lower scores on achievement tests than on aptitude tests. However, given that both aptitude and achievement tests often assess the same underlying knowledge and skills, the choice of labeling one test or another an aptitude or achievement test is generally arbitrary. That means that one could just as easily assert that individuals have higher or lower aptitude in association with their achievement test performance, which makes little conceptual sense but is entirely consistent with the underlying properties of the tests. In fact, given the nature of statistical regression-to-the-mean phenomena, which are associated with taking the difference between any two measures, it is common for individuals with low scores on one test (e.g., aptitude) to have relatively higher scores on the other test (e.g., achievement), and similarly, individuals with higher-than-average scores on one test will have somewhat lower scores on the other test. The attribution that low-aptitude individuals are often overachievers and high-aptitude individuals are often underachievers is most often an artifact of this regression-to-the-mean phenomenon and thus does not provide any useful diagnostic information. Only extremely large differences between such scores (i.e., differences that significantly exceed the difference attributable to

regression-to-the-mean effects) can provide any potential diagnostic information.

—Phillip L. Ackerman

See also Ability Tests; Armed Services Vocational Aptitude Battery; Differential Aptitude Test; Multidimensional Aptitude Battery

Further Reading

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice Hall.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Thorndike, R. L. (1963). *The concepts of over- and under-achievement*. New York: Bureau of Publications, Teachers College, Columbia University.

Scholastic Aptitude Test information: <http://www.college-board.com>

Test reviews: <http://www.unl.edu/buros/>

Testing standards and procedures information: <http://www.apa.org/science/testing.html>

AREA CHART

Area charts are used to simultaneously display visual information in multiple categories. The categories are stacked so that individual as well as cumulative values are shown.

Area charts, like other charts, such as bar and column charts, are most often used for categorical data that are, by definition, not dynamic in nature. For example, if one were interested in illustrating money spent across the first quarter of the year in certain categories, an area chart would be a useful way to do so, as in the following example. First, here are the data:

	<i>Food</i>	<i>Car</i>	<i>Fun</i>	<i>Miscellaneous</i>
January	\$165	\$56	\$56	\$54
February	\$210	\$121	\$87	\$34
March	\$227	\$76	\$77	\$65

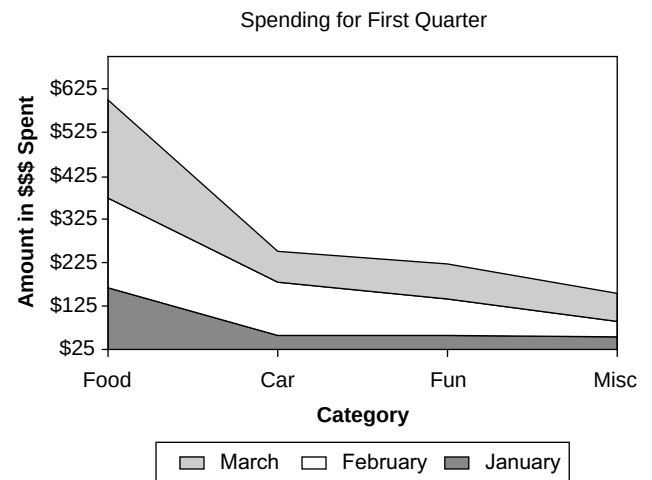


Figure 1 Area Chart Created in Excel

These data represent the amount of money spent in four categories, food, car, fun, and miscellaneous, in each of the first three months of the year. The value for each category by month is shown in the area chart below, with each month containing a series of data points.

An area chart is not an easy chart to create manually; Figure 1 was created using Excel.

—Neil J. Salkind

See also Bar Chart; Line Chart; Pie Chart

Further Reading

Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

ARITHMETIC MEAN

The most widely used measure of central tendency is the arithmetic mean. Most commonly, *mean* refers to the arithmetic mean. The arithmetic mean is defined as all the scores for a variable added together and then divided by the number of observations. Therefore, the formula to compute the arithmetic mean is as follows:

$$\bar{X} = \frac{\Sigma X}{n},$$

where

X represents the data points,

Σ is the summation of all the X s,

n is the number of data points or observations, and

$\bar{X}$ is the computed mean.

For example, take the data presented in Table 1.

The sum of the observations (ΣX) is $1 + 5 + 7 + 2 + 10 + 4 + 6 + 5 + 4 + 6 = 50$. Then, we divide this value by n , which in this example is 10 because we have 10 observations. Thus, $50/10 = 5$. The arithmetic mean for this set of observations is 5.

The SPSS statistical software package provides several ways to compute a mean for a variable. The Mean command can be found under Descriptives, then Frequencies, Explore, and finally Means. Furthermore, the mean can be added to output for more advanced calculations, such as multiple regression. The output for the Descriptives mean is presented in Figure 1.

As seen in the output, the variable “data points” has a total of 10 observations (seen under the column headed N), the lowest value in the data set is 1, the highest value is 10, the mean is 5, and the standard deviation is 2.539.

There are two major issues you should be aware of when using the arithmetic mean. The first is that the arithmetic mean can be influenced by outliers, or data values that are outside the range of the majority of the

Descriptive Statistics

	<i>N</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Mean</i>	<i>Std. Deviation</i>
Data points	10	1	10	5.00	2.539
Valid N (listwise)	10				

Figure 1 SPSS Output for the Descriptives Mean

data points. Outliers can pull the mean toward themselves. For example, if the data set in Figure 1 included a data point (which would be observation 11) of 40, the mean would be 8.2. Thus, when the data set is extremely skewed, it can be more meaningful to use other measures of central tendency (e.g., the median or the mode).

The second issue is that the arithmetic mean is difficult to interpret when the variable of interest is nominal with two levels (e.g., gender) and not meaningful when there are more than two levels or groups for a given variable (e.g., ethnicity). The mean has been found to be consistent across time. With repeated measures of the same variable, the arithmetic mean tends not to change radically (as long as there are no extreme outliers in the data set). Furthermore, the arithmetic mean is the most commonly used measure of central tendency in more advanced statistical formulas.

—Nancy L. Leech,
Anthony J. Onwuegbuzie, and Larry Daniel

See also Average; Median

Further Reading

- Glass, G. V., & Hopkins, K. D. (1995). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.

Table 1 Data for Arithmetic Mean Computation

<i>Observation</i>	<i>Data Points</i>
1	1
2	5
3	7
4	2
5	10
6	4
7	6
8	5
9	4
10	6

ARMED FORCES QUALIFICATION TEST

The Armed Forces Qualification Test (AFQT) is the name given to a series of tests constantly in use from 1950 to the present. The current AFQT is a composite

of subtests from the Armed Services Vocational Aptitude Battery (ASVAB). Constituent members of the composite are verbal and quantitative components, weighted to be of equal variance contribution. The AFQT is used by the U.S. military as a measure of quality and trainability. Although AFQT scores are reported in percentiles, it also has five categories that are used for legal and classification purposes. By law, the American military is not permitted to enlist individuals who score in the lowest category, that is, at the 10th percentile or less, except by direction or legal exception.

The AFQT is the successor to the World War II Army General Classification Test and the Navy General Classification Test. Although these two tests had similar names, their content was not identical. The first AFQT (1950) was equated to a mixed Army-Navy sample based on percentile equivalent scores from these two tests. The American military calls this procedure “calibrating” and used it for many years. All forms of the AFQT were scored on the basis of a normative sample of men in uniform as of 1944.

The 1950 AFQT had four content areas; verbal, arithmetic, spatial, and spatial visualization. In 1953, the content areas of the AFQT were changed to verbal, arithmetic, spatial, mechanical ability, and tool knowledge. The content remained more or less unchanged through the first seven versions of the test. When the ASVAB was adopted (1976), the AFQT became a part of it. In ASVAB versions 5, 6, and 7, the AFQT strongly resembled the numbered AFQT forms. Scores were reported on the 1944 normative sample.

With the implementation of ASVAB forms 8, 9, and 10, the AFQT portion was reworked to add a timed subtest, Numerical Operations. With the implementation of the 1980 nationally representative normative sample in 1984, a problem was encountered with the timed subtests of the ASVAB. A small mistiming of these subtests could and did have large impacts on the resultant scores. The AFQT composite was then changed to be equally verbal and arithmetic in importance. Scores were still reported on the metric of the 1980 normative sample.

In 1997, another nationally representative normative sample was collected, and the ASVAB subtests,

including those contributing to the AFQT, were placed on this new metric as of July 1, 2004. The current AFQT still consists of verbal and arithmetic components. It is basically a measure of general cognitive ability.

Although the AFQT is not available for commercial use, it has been used in nonmilitary research, most notably in Herrnstein and Murray’s *The Bell Curve* (1994) and in numerous econometric and occupational studies. It has been offered as an ability measure and related to training and job performance, earnings, educational attainment, and interregional migration.

—Malcolm James Ree

See also Ability Tests

Further Reading

Orme, D. R., Brehm, W., & Ree, M. J. (2001). Armed Forces Qualification Test as a measure of premorbid intelligence. *Military Psychology, 13*(4), 187–197.

ARMED SERVICES VOCATIONAL APTITUDE BATTERY

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude battery used for two purposes. The first is proprietary: enlistment qualification and job classification for all the branches of the American military and the Coast Guard. The second purpose is to provide vocational guidance for high school and vocational students. These two uses are called the Enlisted Testing Program (ETP) and the Student Testing Program (STP), respectively.

The ASVAB consists of nine subtests. In the past, it had contained some purely timed subtests and some subtests well characterized as mixed speed and power. Subtest content is verbal ability, quantitative ability, spatial ability, and job knowledge. The subtests are General Science, Arithmetic Reasoning, Mathematics Knowledge, Word Knowledge, Paragraph Comprehension, Electronics Information, Auto and Shop Information, Mechanical Comprehension, and

Assembling Objects. Paragraph Comprehension is a short subtest and is never used alone but always combined with Word Knowledge to increase reliability. All subtest raw scores are converted to normative population standard scores and then combined to make composites for use.

For use in the ETP, the ASVAB is administered in both paper-and-pencil and computer-adapted forms. In the ETP, all armed service members are required to pass a composite called the Armed Forces Qualification Test (AFQT) for minimal enlistment qualification. In addition to the AFQT, each service computes unit or simple weighted composites for classification of enlistees into specific occupations or clusters of occupations. The scores are reported either in normative percentiles or in service-specific normative standard scores. In the normative sample, the composites reflect the high loading of general cognitive ability in all the subtests. This loading also makes the composites valid for predicting occupational criteria.

The STP uses only paper-and-pencil administration and is aimed at career exploration. The ASVAB and its supporting interpretive materials are offered free of charge. In this program, the scores are reported by grade norms in the form of percentiles with error bands. Grades are reported for each of eight subtests and three content composites. The STP does not use the Assembling Objects subtest. Included in the program are copious materials to aid the student in exploring potential occupational goals. In the STP, an associated interest inventory is offered, and the relationships between interests and test scores are clearly explained in a guide titled *Exploring Careers*.

—Malcolm James Ree

See also Armed Forces Qualification Test

Further Reading

Exploring Careers: The ASVAB career exploration guide. (2005). DD Form 1304-5WB, July 2005. U.S. Government Printing Office.

Ree, M. J., & Carretta, T. R. (1994). Factor analysis of ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement*, 54, 457–461.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology*, 44, 321–332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance; Not much more than g. *Journal of Applied Psychology*, 79, 518–524.

ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a class of models, inspired by the central nervous system, used in machine learning and pattern recognition and classification. These models are nonlinear parametric regression models with either automatic, unsupervised training (setting of the model parameters) or supervised training from some training set of known input-output relations, depending on the type of network.

An ANN consists of a collection of *neurons* (processing units) and connections between these neurons. Usually these neurons accumulate “information” from neighboring neurons and either fire or not, depending on some local threshold level.

The simplest type of ANN is the *feed-forward net*, in which all the information flows in a single direction. Figure 1 shows a four-layer feed-forward net, with an *input* layer, followed by two *hidden* layers, followed in turn by an *output* layer.

Many neurons are modeled (and constructed or simulated) to have binary outputs (and often binary inputs as well). Each neuron has some rule (called a

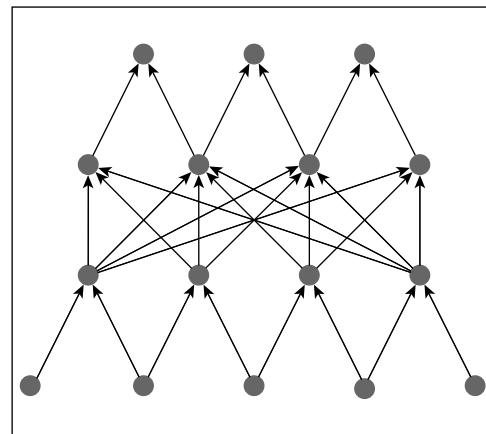


Figure 1 A Simple Four-Layer Feed-Forward Net

firing rule) for deciding which combination of inputs results in which output. One particularly simple rule would be to sum all the inputs (multiplied perhaps by weights) and check to see if the sum was more than some threshold value; if so, then fire, and if not, then do not fire (a binary output). Notice that this firing rule is discontinuous in that it has a sudden jump at the threshold. This rule is sometimes called the *hardlimit* firing rule.

As a simple example (for the neuron illustrated in Figure 2), let $w_1 = 0.3$, $w_2 = 0.5$, $w_3 = 0.3$, and the threshold $q = 1.5$. Then for the inputs 1.2, 2.1, and 0.7, we get accumulated sum $(.3)(1.2) + (.5)(2.1) + (.3)(0.7) = 1.62 > 1.5$, so the neuron would fire.

The parameters of a feed-forward net include the weights and thresholds for each neuron. These parameters must be set during the *training phase* of the network. During training (a *supervised learning* situation), one uses a set of known input-output relationships to set the parameters of the network. One common technique for training is the *backpropagation* algorithm. This algorithm basically computes a gradient of the error with respect to the weights in order to adjust the weights. The computation proceeds by propagating influences backwards in the network (and hence the name).

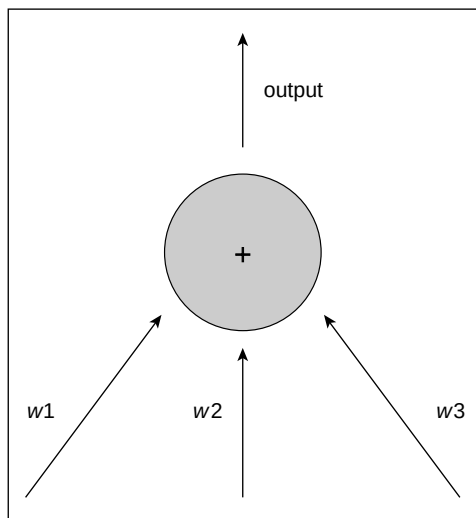


Figure 2 A Single Neuron

Many other kinds of firing rules are possible, including a sigmoid (a continuous version of the

hardlimit rule), Gaussian rules, and others. The sigmoid firing rule has the functional form

$$\text{output} = \frac{1}{1 + \exp(-\text{sum of weighted inputs}/T)}.$$

Many other kinds of connection *topologies* are possible. A *Hopfield* net, for example, is a neural network with bidirectional connections (and with the same weight in each direction). These networks act as associative memories in the sense that the network can store a set of patterns and fire if a similar pattern is presented to it.

The presence of loops in a network allows for feed-back, so these types of networks are sometimes called *recurrent* networks.

Many popular software systems, including Matlab, SPSS, SAS, and R (a open source version of S), have neural network toolboxes. It is also fairly easy to get software packages (as either executables or source) from the Internet to simulate various neural networks; one such place is <http://www.neural-networks-at-your-fingertips.com/>.

—Franklin Mendivil

Further Reading

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley.

Artificial neural networks technology: http://www.dacs.dtic.mil/techs/neural/neural_ToC.html

ASSESSMENT OF INTERACTIONS IN MULTIPLE REGRESSION

With theories growing ever more complex, analytic tools for their analysis and testing need to be developed. In exploring moderator variables that are found in theory tests in both experimental and nonexperimental research, we need to be careful to assess the interactions between the moderator variable and the predictor

variable(s) in an appropriate way. Prior to the early 1990s, many nonexperimentalists using a correlational paradigm often used it inappropriately by (a) correctly creating a product term for the moderator and independent variable and then (b) inappropriately correlating it with the dependent variable. This paradigm is inappropriate for both theoretical and empirical reasons.

Theoretically, as Jacob Cohen has argued, while the interaction is carried by the product term, it is not the product term. The product term alone also carries variance due to the main effects of the independent variable and the moderator variable. The appropriate analysis is to partial out the main effects in a multiple regression analysis, as pointed out by Saunders as long ago as 1956.

Empirically, as Schmidt has shown, the correlation between a product term and the dependent variable is sensitive to the scale numbers used in the analysis. Changing from a scale of 1 to 5 to a scale of -2 to +2 will change the correlation dramatically. The proper analysis, as Arnold and Evans have shown, results in the incremental R^2 between an equation containing just main effects and one containing the main effects plus the product term being invariant under linear transformations of the data (unlike the simple correlation between the product term and the dependent variable, which changes dramatically). This invariance translates to a proper test of theory only if the measurement scales and the underlying psychological constructs are linearly related. More recent developments involving structural equation modeling do not have this limitation.

—Martin G. Evans

See also Analysis of Variance (ANOVA); Correlation Coefficient

Further Reading

- Arnold, H. J., & Evans, M. G. (1979). Testing multiplicative models does *not* require ration scales. *Organizational Behavior and Human Performance*, 24, 214–224.
- Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equations models. *Structural Equation Modeling*, 5, 267–293.
- Busemeyer, J., & Jones, L. R. (1983). Analysis of multiplicative causal rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562.

- Cohen, J. (1978). Partialled products *are* interactions; partialled powers *are* curve components. *Psychological Bulletin*, 85, 858–866.
- Evans, M. G. (1985). A Monte-Carlo study of correlated error in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes*, 36, 305–323.
- Evans, M. G. (1991). The problem of analyzing multiplicative composites: Interactions revisited. *American Psychologist*, 46, 6–15.
- Ping, R. A. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166–175.
- Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement*, 16, 209–222.
- Schmidt, F. L. (1973). Implications of a measurement problem for expectancy theory research. *Organizational Behavior and Human Decision Processes*, 10, 243–251.

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

The Association for Psychological Science (APS) is the leading national organization devoted solely to scientific psychology. Its mission is to promote, protect, and advance the interests of scientifically oriented psychology in research, application, and improvement of human welfare.

Established in 1988, APS was instantly embraced by psychology's scientific community, and its membership grew rapidly. By the end of its first year, APS opened an office in Washington, D.C., and now has approximately 15,000 members from around the world. Members are engaged in scientific research or the application of scientifically grounded research spanning all areas of psychology. There are also student affiliates and institutional members. Distinguished contributions are recognized by Fellow status.

Formation

APS was created out of recognition that (a) the needs and interests of scientific and academic psychologists are distinct from those of members of the professional

community primarily engaged in clinical practice and (b) there was a strong need for a society that would advance the interests of the discipline in ways that more specialized organizations were not intended to do. An interim group, the Assembly for Scientific and Applied Psychology, had sought to reform the American Psychological Association from within, but its efforts were rejected by a membership-wide vote of the APA. APS then became the official embodiment of the reform effort, and the new organization was launched on August 12, 1988.

Publications

APS publishes four journals:

1. *Psychological Science* publishes authoritative articles of interest across all of scientific psychology's subdisciplines.
2. *Current Directions in Psychological Science* offers concise invited reviews spanning all of scientific psychology and its applications.
3. *Psychological Science in the Public Interest* provides definitive assessments by panels of distinguished researchers on topics for which psychological science has the potential to inform and improve the well-being of society.
4. APS's newest journal, *Perspectives on Psychological Science*, features longer integrative reviews and a variety of eclectic articles.

APS also publishes the monthly *Observer*, featuring news and opinion pieces; a Current Directions Readers series in conjunction with Prentice Hall; a Festschrift series in conjunction with LEA Press; and self-published books on the teaching of psychology.

Annual Convention

APS holds a meeting in late spring each year to showcase the best of scientific psychology. The program features presentations by the field's most distinguished researchers and educators in a variety of formats, including invited addresses and symposia, submitted symposia, "hot topic" talks, and posters. The convention also includes workshops on specialized topics.

APS Fund for the Teaching and Public Understanding of Psychological Science

In 2004, the David and Carol Myers Foundation pledged \$1 million to APS for the creation of an endowed fund that aims "to enhance the teaching and public understanding of psychological science for students and the lay public, in the United States, Canada, and worldwide."

Achievement Awards

APS recognizes exceptional contributions to scientific psychology with two annual awards: the APS William James Fellow Award (for significant intellectual contributions to the basic science of psychology) and the James McKeen Cattell Fellow Award (for outstanding contributions to the area of applied psychological research).

APS Student Caucus

Students are an important and active component of APS. The APS Student Caucus is the representative body of the society's student affiliates. The Student Caucus organizes research competitions, convention programs, and a variety of membership activities aimed at professional development and enhanced education in psychological science.

Advocacy

APS is widely recognized as an active and effective leader in advancing the interests of basic and applied psychological, behavioral, and social science research in the legislative arena and in the federal agencies that support these areas of research.

—Morton A. Gernsbacher,
Robert W. Levenson, and Sarah Brookhart

See American Psychological Association

Further Reading

Association for Psychological Science Web site: www.psychologicalscience.org

ASYMMETRY OF *g*

Is the strength of the general factor of intelligence uniform? That is, do the various indicators of crystallized intelligence correlate similarly at all levels of intelligence? This question was first explored in 1927, by Spearman, who, on the basis of his data, suggested a law of diminishing returns (also known as the divergence hypothesis, or in my terms, the asymmetry of *g*). This observation states that at low levels of intelligence, the various facets of intelligence are very highly correlated, but at higher levels of intelligence, the correlations between the various facets are less strong.

Between 1927 and 1990, this work disappeared from sight, and as late as 1989, Detterman and Daniel could argue that “it was thought that the positive manifold [the loading of ability tests onto a single factor] was uniformly distributed over the whole range of ability.” This finding has been supported in a number of other studies.

The best estimate is for a linear decline in average correlations between IQ measures from $r = 0.46$ for the least gifted group to average correlations of about $r = 0.30$ for the more highly gifted. If this picture is sustained through additional research, then perhaps much of the conflict between those arguing for a pervasive *g* factor and those arguing for specific abilities can be reconciled by considering that both may be true, with *g* dominating at the lower levels of intelligence and specific factors dominating at higher levels of intelligence.

In practice, this means that we will have to reconsider our use of intelligence as a selection tool in job performance. It is fine to use *g* at low levels of intelligence, but at high levels, the specific abilities needed for a task will come into play, so predictive studies will have to explore the interaction between *g* and specific abilities to understand who the high performers will be.

—Martin G. Evans

See also Ability Tests; Intelligence Tests

Further Reading

- Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence*, 23, 105–132.
- Detterman, D. K., & Daniel, M. H. (1989). Correlates of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, 13, 349–359.
- Evans, M. G. (1999). On the asymmetry of *g*. *Psychological Reports*, 85, 1059–1069.
- Evans, M. G. (2002). The implications of the asymmetry of *g* for predictive validity. In W. Auer-Rizzi, E. Szabo, & C. Innreiter-Moser (Eds.), *Management in einer Welt der Globalisierung und Diversitaet: Europaeische und nordamerikanische Sichtweisen* (pp. 433–441). Stuttgart, Germany: Schaeffer-Poeschel Verlag.
- Jensen, A. R. (2003). Regularities in Spearman's law of diminishing returns. *Intelligence*, 31, 95–105.
- Spearman, C. E. (1927). *The abilities of man*. London: Macmillan.

ATTENUATION, CORRECTION FOR

Attenuation is a term used to describe the reduction in the magnitude of the correlation between the scores of two measurement instruments that is caused by their unreliabilities. Charles Spearman first recognized the value of correcting for attenuation by noting that we are interested in determining the true relationship between the constructs we study, not the relationship between flawed empirical measures of these constructs. His solution was to estimate the correlation between two variables using perfectly reliable empirical measures. He developed the following formula:

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}, \quad (1)$$

where r_{xx} and r_{yy} equal the reliability coefficients of the two instruments, and r_{xy} equals the obtained correlation between the scores of the two instruments. It is assumed that X and Y are imperfect measures of underlying constructs X' and Y' , containing independent, random errors, and $r_{x'y'}$ equals the true correlation between X' and Y' . If r_{xx} equals .61, r_{yy} equals .55, and r_{xy} equals .43, then

$$r_{x'y'} = \frac{.43}{\sqrt{(.61)(.55)}} = .74. \quad (2)$$

The use of the formula allows the researcher to answer the question, What would the correlation (i.e., $r_{x'y'}$) be if both of the empirical measures were error free? The example illustrates the considerable reduction in the size of r_{xy} caused by the unreliabilities of the scores for the two instruments.

Although correcting for the attenuation of both empirically measured constructs is useful for investigating theoretical problems, it is more difficult to justify for practical applications. For instance, when predicting who will succeed in higher-education programs or who will benefit from special education services, we are limited by the fallible instruments at hand; after all, the application of the correction for attenuation does not make the empirical scores more reliable than they really are! Although it may not be appropriate to correct for attenuation of a predictor variable, it may well be justifiable to adjust for the inaccuracy of criterion measures. For instance, why should inaccurate graduate grade point averages be allowed to make Graduate Record Examinations scores appear less valid than they really are? For this single correction problem, the formula is as follows:

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{yy}}}. \quad (3)$$

Confirmatory factor analysis (CFA) provides a second way to correct for attenuation. In CFA, the measurement error of each latent variable is explicitly modeled. In research comparing the results of correcting for attenuation via the two approaches, Fan found highly comparable results for the same data. That is, both approaches provided nearly identical point estimates and confidence intervals for the relationship between the true scores of his variables. Nevertheless, it should be mentioned that the CFA approach might be less applicable, given the constraints of modeling item-level data (e.g., extreme item skewness and kurtosis, different item distributions, and item unreliability).

History

Over the years, various writers have debated whether attenuation should be corrected at all. Although he supported the use of correction for attenuation, Nunnally called it a “bandy fiction” the results of which are always hypothetical. However, given its use in adjusting effect sizes in substantive research and meta-analysis, it appears that correction for attenuation is here to stay. One reason is the alarm expressed by some scholars that doctoral programs in the United States are giving short shrift to the measurement curriculum. Some suggest that the lack of attention given measurement issues in higher education has led to the finding that as many as 50% of contemporary published research articles fail to report the reliability and validity of the independent and dependent variables employed. This fact takes on added importance when one realizes that some variables used in published research are so unreliable as to make it virtually impossible to obtain statistically significant results. Furthermore, an increasing number of professional journals have begun to suggest strongly, and even demand, that appropriate corrections be provided readers in order to better inform their judgments regarding the practical importance of statistically significant p values.

Which Type of Reliability Should Be Used?

One pertinent issue concerns the selection of the type of reliability coefficient that should be used: test-retest, internal consistency, or alternative-form reliability. Obviously, the selection of the type of reliability influences the corrected true-score correlation. For instance, the long-term stability of an instrument (e.g., using a 1-year test-retest interval) is expected to be much lower than an internal consistency estimate of reliability. Consider the case in which the 1-year test-retest reliability coefficient for a criterion test equals .65, coefficient alpha (an estimate of internal consistency reliability) equals .80 for the same test, and the validity coefficient between the predictor variable and the criterion test is .49. If we apply

Equation 2 to correct for the attenuation of the criterion variable on the validity coefficient, we obtain $r_{X'Y'}$ equal to .61 for the test-retest estimate and .55 for the coefficient alpha estimate. Without knowing the specific interest of the researcher, we have no basis for identifying the better estimate of the true validity coefficient. Over the years, different writers have suggested aggregating corrections for attenuation across several data sets and using the average value as the best estimate of the true relationship between the sets of scores. Cronbach and others have pointed out that, should the average value be based on corrections using different sorts of reliability coefficients, a basic requirement of classic measurement theory would be violated: Different types of reliability estimates may not be used interchangeably.

In the literature, each of the three forms of reliability has been recommended for use in correction for attenuation. Most recently, researchers are advised to choose on the basis of the intent and context of their research objectives. Although such advice is easy to follow in some contexts, it is murky in others. For instance, if one is interested in determining the ability of a first-grade readiness test to predict student academic achievement 1 year later, then the test-retest reliability (i.e., 1-year stability) of the readiness test would be of greater interest to the researcher than would other forms of reliability. Should the researcher be interested in increasing the number of items in a test (and thus improving the sampling of the item domain and increasing reliability) in order to increase the validity of the instrument, then internal consistency reliability estimates offer more informative results.

What Is Done When Appropriate Reliability Estimates Are Unavailable

Although it is less problematical in individual research reports, in meta-analyses it often happens that reliability estimates of the appropriate types are unavailable to the analyst. There are no elegant solutions to this dilemma; the following are some of the common practices in calculating the correction for attenuation:

1. When appropriate reliability coefficients are presented in the published report, they are used to correct for attenuation.
2. When only inappropriate reliability coefficients are presented in the published report, they are used to correct for attenuation.
3. When no reliability coefficients are presented in the published report, published values from technical manuals and other sources are used to correct for attenuation.
4. When no published reliability coefficients are available for use, then average reliability coefficients from other, similar instruments are used to correct for attenuation.

Clearly, this is not an enviable state of affairs. However, until such time as researchers and professional publications rectify the omissions that remain commonplace in published reports, the correction for attenuation will remain a stepchild in research practice.

—Ronald C. Eaves

See also Correlation Coefficient

Further Reading

- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, 62, 254–263.
- Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, 63, 915–930.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127, 162–181.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1, 199–223.

- Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. (Available for purchase at <http://www.psycinfo.com/psycarticles/1999-03403-008.html>)
- Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: A meta-analytic generalization across studies. *Educational and Psychological Measurement*, 60, 201–223.

ATTITUDE TESTS

Although definitions have varied, the most common contemporary definition is that an attitude is a relatively general and enduring evaluation of a person, object, or concept along a dimension ranging from positive to negative. Because attitudes have played a central role in psychological theory and application, numerous and diverse attitude tests have been developed to assess this construct.

Direct Measures of Attitudes

The most common way to assess attitudes is simply to directly ask people to report them. Traditionally, this approach has been accomplished using formal scaling procedures to construct multiple-item tests of attitudes.

Multiple-Item Direct Attitude Measures

Formal attitude measurement began with the work of Louis L. Thurstone. Although he proposed several procedures, his most popular approach was the equally appearing interval scale. This procedure requires development of a series of opinion statements that represent varying levels of positivity or negativity toward the attitude object. Judges then sort these statements into equal-interval categories of favorability toward the attitude object (e.g., 11 intervals, where 1 = *extremely unfavorable* and 11 = *extremely favorable*). Next, a scale value is computed for each statement that corresponds to the median (or mean) score of judges' ratings, eliminating items with highly

variable ratings. From the remaining items, the researcher selects a final set of statements representing equal intervals on the evaluative dimension. The final scale consists of these items in random order, with instructions for respondents to check the items with which they personally agree. The mean scale values of marked items are computed to obtain individual attitude scores.

Although Thurstone scales generally work well, the procedure is time consuming because it requires items to initially be rated by judges. In response to this and other concerns, in 1932 Rensis Likert developed the method of summated ratings. This procedure requires a set of opinion items that are clearly positive or negative in relation to the attitude object. The items are then administered to the sample of interest, whose members are instructed to indicate their level of agreement on a 5-point continuum (where *strongly agree* is assigned a value of 5 and *strongly disagree* is represented by 1). When the sample has completed the items, negative items are reverse coded, and each respondent's item scores are summed to create an overall attitude score. Item-total correlations are computed to identify poorly performing items (i.e., items with low item-total correlations). These items are discarded, and the final attitude scale scores are computed. Research has suggested this procedure tends to produce highly reliable scales.

In response to concerns that prior procedures did not guarantee unidimensional attitude tests, Louis Guttman proposed scalogram analysis. This approach involves constructing a set of opinion statements that are ranked in order, from least extreme to most extreme (i.e., a set of statements ranging from mildly positive to extremely positive or a set of items ranging from mildly negative to extremely negative). Scalogram analysis assumes that agreeing with a more extreme position implies agreement with less extreme positions. The pattern of responses to the set of items is examined to assess the extent to which items satisfy this assumption. Items are discarded if they frequently fail to meet this criterion. Although Guttman scales have the advantage of producing scales that are likely to be unidimensional, the difficulty of constructing scales that satisfy its stringent requirements has

prevented it from becoming a widely used method of attitude measurement.

One of the most popular measures of attitudes is Charles Osgood's semantic differential. The semantic differential consists of a set of bipolar adjective pairs (e.g., good-bad, wise-foolish) that represents the evaluative continuum. Because these bipolar adjective pairs tend to be very general, they can typically be applied to any attitude object. Usually the attitude object is placed at the top of the page and participants are asked to rate this object by marking a category on each of the bipolar pairs (reserving the middle category for neutral responses). Attitudes are calculated by summing or averaging the scores for each bipolar scale. Semantic differential scales have generally been found to have adequate psychometric properties and are easy to use. However, because of their generality, they may provide somewhat less precise and comprehensive measurement of specific attitudes than Thurstone or Likert scales do.

Single-Item Direct Attitude Measures

Attitude measures are traditionally composed of multiple items. However, in some cases, attitudes are measured with a single semantic differential item or opinion statement. This practice is common in situations such as telephone surveys in which attitudes toward a wide range of topics must be assessed in a short time. Unfortunately, any single item may have biases or ambiguities and may lack the breadth to fully capture an attitude. Thus, single-item measures can have inadequate reliability and validity. A large literature has developed regarding the optimal way to construct such items and organize them within a questionnaire.

Alternatives to Direct Measures of Attitudes

Direct measures of attitudes presume people are willing and able to discuss their attitudes openly. This may not always be the case, and researchers have developed numerous alternatives to direct methods of attitude measurement.

Indirect Measures of Attitudes

One of the earliest and best known indirect measures is the information-error test, developed by Kenneth Hammond. This procedure begins with generating a large number of objective knowledge questions about the target issue. The goal is to create questions that in principle have objective answers but to which respondents are unlikely to know the answers. These questions are presented in a multiple-choice format, with answers implying various levels of negativity or positivity toward the target position (e.g., a question on capital punishment might ask respondents to guess the percentage of executed criminals who were later found to have been falsely convicted). The assumption underlying this method is that when individuals are faced with questions for which they do not know the answer, they will tend to guess in an attitude-consistent manner. Attitude scores can be calculated by assigning scale values to increasingly positive or negative answers and then summing or averaging the responses across items.

More recently, indirect attitude measures have received renewed attention stemming largely from the increasing literature on unconscious (or implicit) attitudes and the suggestion that indirect measures may be more effective than direct measures in capturing unconscious attitudes. As a result, new indirect measures (called implicit measures of attitudes) are being introduced with increasing frequency. Two implicit measures have received particular attention.

Russell Fazio's evaluative priming technique (also called the bona fide pipeline or affective priming) is based on the notion that even if we are not specifically asked to report our attitudes, they will automatically come to mind when we encounter the attitude object. Interestingly, automatically activated attitudes influence our judgments of other objects. It has been shown that if a negative attitude is activated, people can make quick judgments about other objects that are negatively evaluated but are slower in making judgments about objects that are positively evaluated. The opposite is true if the initial attitude activated is positive. This effect is central to the evaluative priming technique.

The evaluative priming procedure involves presenting respondents with a dual judgment task. They are told that first they will be briefly presented on computer with a word prime that they should try to remember and then they will be presented with a second word, for which they will need to make an evaluative judgment (i.e., judge if the word is positive or negative). Respondents are presented with numerous word pairs, and the time it takes them to make judgments about the second word in the pair is recorded. Attitudes are measured by including the target attitude object among the primes. The speed of judgments using the attitude object are then examined when it serves as a prime for words that are almost universally liked (e.g., *love*) and as a prime for some words that are almost universally disliked (e.g., *death*). If a person's attitude is positive, that person should be faster in making judgments about the second words when the attitude object is a prime for positive words than when it is a prime for negative words. The reverse is true for negative attitudes.

Another recently developed indirect test is the Implicit Association Test (IAT). This measure is designed to tap the (automatic) associations between two concepts (e.g., homosexual, heterosexual) with positive and negative evaluations. In this procedure, respondents are told that they will be given a list of words on computer that will fall into one of four categories: the target attitude object (e.g., *homosexuals*), a comparison attitude object (e.g., *heterosexuals*), positive words, and negative words. Respondents assign each word to one of the four categories by pressing one of two response keys. In one set of trials, respondents are told to hit one key if the word is a word related to the target object *or* a positive word and to hit a different key if the word is related to the comparison object *or* a negative word. In a second set of trials, the task is reversed so that target object words and negative words share the same response key, and comparison object words and positive words share the same response key. The computer records the time it takes for respondents to make their categorizations.

The logic underlying the IAT is that if a person's attitude toward the target attitude object is positive, that person will be able to perform the task more

quickly when target object words share the same response key with positive words than they will when target object words share the same response key with negative words. In contrast, if the person's attitude is negative, the reverse will be true. Attitude scores are created by computing a numerical index that reflects the difference between the average speed with which people perform the two versions of the task.

Both the evaluative priming procedure and the IAT are new measures of attitudes, and research is still being conducted to assess their reliability and validity.

Physiological Measures of Attitudes

Another alternative to direct measures of attitudes is the use of physiological measures. Although some early physiological measures of attitudes were problematic, more recent measures have proved promising. For example, the use of event related brain potentials (ERP) has proven effective in assessing attitudes. The ERP is measured by attaching electrodes to certain areas of the scalp and monitoring the pattern of electrocortical activity that occurs when people are categorizing objects. Specifically, a sequence of stimuli that is evaluatively consistent (i.e., all positive or all negative) is presented to the participant, with the target attitude object at the end of this sequence. If the attitude object differs in categorization from the previous stimuli (e.g., it is evaluated negatively and the previous sequence is positive), a large ERP will occur. If the attitude object's categorization is consistent with the previous objects (e.g., all positive), a small ERP will occur. By presenting the attitude object at the end of both a sequence of positive objects and a sequence of negative objects, the overall attitude score can be computed by comparing the size of the ERP for the two sequences.

Recently, brain-imaging techniques have been applied to the measurement of attitudes. For example, functional magnetic resonance imaging (fMRI), a procedure for measuring changes in brain activity through increases in blood flow and oxygen consumption, can be used to assess attitudes. By placing participants in the fMRI scanner and presenting them with an image of the attitudinal object, researchers can see

what parts of the brain are activated in response to the target stimuli. Such activation can then be examined in relation to distinct patterns of brain activity associated with particular emotional and cognitive responses. Use of this technique enables researchers to assess attitude valence and possibly even attitude structure (e.g., emotional or cognitive).

—Sonia Matwin and Leandre R. Fabrigar

See also Guttman Scaling; Likert Scaling; Questionnaires; Thurstone Scales

Further Reading

- Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
- Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In C. T. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 17–40). Thousand Oaks, CA: Sage.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken, *The psychology of attitudes* (pp. 23–87). Orlando, FL: Harcourt Brace Jovanovich.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Project Implicit: <https://implicit.harvard.edu/implicit/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Shlaes, J. L., Jason, L. A., & Ferrari, J. R. (1999). The development of the Chronic Fatigue Syndrome Attitudes Test: A psychometric analysis. *Evaluation & the Health Professions*, 22(4), 442–465.

Chronic Fatigue Syndrome (CFS) is characterized by debilitating symptoms including persistent or relapsing fatigue, and as a result of CFS, some individuals experience significant stigma. Many medical professionals are skeptical of the validity of the illness, and employers often fail to appreciate its seriousness. There is presently no tool to measure attitudes toward this illness or toward people who have CFS. The purpose of these studies was to create a scale that measures attitudes toward individuals with CFS—the Chronic Fatigue **Attitudes Test** (CAT)—and to assess the scale’s reliability and validity. The 13-item scale was created using several constructs outlined in the literature regarding negative attitudes toward people with CFS, disabilities, and AIDS.

ATTRIBUTABLE RISK

The attributable risk statistic provides an estimate of the proportion or number of events that can be explained by a particular risk factor. Epidemiologists frequently use attributable risk calculations to determine the population impact associated with a disease,

Table 1 Attributable Risk Hypothetical Sample

		Lung cancer at follow-up		
		Yes	No	
Event	No event			
Exposed	A	B	A + B	Smoker
Unexposed	C	D	C + D	Nonsmoker
	A + C	B + D		
		3,000	7,000	10,000
		1,000	9,000	10,000
		4,000	16,000	

Notes: Attributable risk = $(A/A + B) - (C/C + D)$; $(3,000/[3,000 + 7,000]) - (1,000/[1,000 + 9,000]) = .30 - .10 = .20$; Attributable risk = $C/(C + D) \times [(A/A + B) / (C/C + D) - 1]$; Attributable risk = $.10 \times (3 - 1) = .20$

behavior, or condition. The U.S. Surgeon General's estimate that smoking accounts for up to 400,000 deaths annually in the United States is an example of an attributable risk inference.

In the context of cohort studies, attributable risk is also referred to as risk difference, in this case quantifying the excess risk in exposed versus unexposed groups. Attributable risk can be calculated in several ways. When accompanied by a relative risk statistic, the attributable risk is equal to the rate of events in the unexposed group $\times$ (relative risk -1). A more widely applicable formula requires the creation of a 2×2 contingency table, as illustrated in Table 1 with a hypothetical sample of smokers and nonsmokers.

In Table 1, A through D represent the number of cases in each study cell. The event rate among smokers (3,000/10,000) is three times higher than among nonsmokers (1,000/10,000). Half the sample is composed of smokers, and 20% develop cancer over the course of the study. Attributable risk uses both the group event rates and the prevalence of the exposure (smoking) for calculation purposes. The attributable risk value of .20 tells us that if smokers in the study became nonsmokers, the incidence of lung cancer would decrease by 20 per 100 individuals. This represents a potential 66% decrease in lung cancer cases.

Important facts about attributable risk:

- The attributable risk statistic alone does not imply that a causal relationship exists between the exposure factor and the event.
- Because attributable risk uses information about exposure prevalence, the attributable risk values between two exposure factors that each double the risk of an event such as cancer can differ dramatically if one exposure (e.g., working in coal mines) is much more rare than a more common exposure (e.g., smoking).
- Attributable risk can be used to calculate the population attributable risk by use of the following formula: attributable risk $\times$ rate of exposure in the population.
- The proportion of events potentially eliminated in a population by changing the exposure rate to that of the unexposed group is often referred to as the attributable proportion.
- By combining attributable risk with the financial costs of a health event, researchers can estimate the

health care expenses associated with an exposure factor and calculate the health care savings achievable by modifying a risk factor in a population.

—Thomas Rutledge

See also Probability Sampling

Further Reading

- Centers for Disease Control and Prevention. (2003, September 5). Cigarette smoking-attributable morbidity—United States, 2000. *MMWR*, 52(35), 842–844. Available from <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5235a4.htm>
- Sedgwick, J. E. C. (2001). Absolute, attributable, and relative risk in the management of coronary heart disease. *Heart*, 85, 491–492.
- Thun, M. J., Apicella, L. F., & Henley, S. J. (2000). Smoking vs other risk factors as the cause of smoking-attributable deaths. *JAMA*, 284, 706–712.

Disease rates comparison: <http://bmj.bmjournals.com/epidem/epid.3.html>

ATTRITION BIAS

When data are collected over two or more points in time, it is common for some participants to drop out of the study prematurely. The attrition of the original sample can occur in longitudinal research as well as in experimental designs that include pretest, posttest, and follow-up data collection. In longitudinal research, which often lasts many years, some participants move between data points and cannot be located. Others, especially older persons, may die or become too incapacitated to continue participation in the study. In clinical treatment studies, there may be barriers to continued participation in the treatment program, such as drug relapse or lack of transportation.

Attrition of the original sample represents a potential threat of bias if those who drop out of the study are systematically different from those who remain in the study. The result is that the remaining sample becomes different from the original sample, resulting in what is

known as attrition bias. However, if sample attrition over time is not systematic, meaning that there are no unique characteristics among those who drop out, then there is no attrition bias, even though the sample has decreased in size between waves of data collection. It is important, then, for researchers who collect multiple waves of data to check for attrition bias.

Attrition bias is one of the major threats to multi-wave studies, and it can bias the sample in two ways. First, attrition bias can affect the external validity of the study. If some groups of people drop out of the study more frequently than others, the subsequent longitudinal sample no longer resembles the original sample in the study. As a result, the remaining sample is not generalizable to the original population that was sampled. For example, a longitudinal sample examining the grieving process of women following the death of a spouse may fail to retain those participants who have become too distraught to fill out the questionnaire. The nonparticipation of this group may bias the findings of the study toward a minimization of depressive symptomatology as a component of the grieving process. In other words, the composition of the sample changes to the point that the results are no longer generalizable to the original population of widows.

Second, systematic, as opposed to random, attrition can negatively affect the internal validity of the study by altering the correlations among the variables in the study. This problem occurs in longitudinal research because the subsamples that are dropping out of the study at a higher rate are underrepresented in the longitudinal sample, which may lead to correlations between variables that are different from the true correlations in the original sample. For example, the underrepresentation of widows with depressive symptomatology in the second or third wave of a study may alter the correlation between insomnia and length of time since the death of the spouse.

Selective attrition affects the internal validity of experimental research when there are differential dropout rates between the treatment and control groups. In a clinical trial of a depression treatment, if the participants in the treatment group drop out at a higher rate than do the participants of the control group, the results of the study will be biased toward

showing artificially successful treatment effects, thus compromising the internal validity of the study. However, if the dropout rates are comparable, the threats to internal validity due to attrition are minimal.

Preventing Attrition

Because of the threat of attrition bias to the external and internal validity of studies, it is important to minimize sample attrition when conducting multiwave research. Researchers who have conducted experimental and longitudinal research have made a number of recommendations and suggestions to reduce sample attrition. Mason emphasized the importance of creating a project identity, offering cash and other incentives, developing a strong tracking system to constantly identify the location and status of participants, and keeping follow-up interviews brief. Others recommend collecting detailed contact information about participants to increase the likelihood of locating them for the second and subsequent interviews. Follow-up postcards and telephone reminders also help retain participants in the sample.

Detecting Attrition Bias

Differences in characteristics between those who prematurely drop out a study (“droppers”) and those who remain in the sample (“stayers”) can be assessed by conducting a logistical regression analysis. Because both groups participated in the first wave of the study, data are available on which to compare the two groups. A dichotomous dependent variable is created with 1 representing the stayers and 0 representing the droppers. Variables from the first wave of data are used as independent variables in the analysis. These variables should include key demographic variables, such as race, income, age, and education, as well as substantive variables that are salient in the study, such as depression, drug abuse, or marital quality. A statistically significant coefficient for any of the variables means that there is a difference between the stayers and the droppers, indicating attrition bias.

Threats to internal validity due to attrition bias can be tested by comparing the first-wave correlation

matrices of the overall sample and the longitudinal sample, which includes only the stayers. This can be done in two ways:

1. Each of the correlation coefficients (for example, the correlation between age and level of depression) is compared using Fisher's z statistical test. A significant z score means that the two coefficients are statistically significantly different, indicating attrition bias.
2. A structural equation modeling program, such as LISREL or AMOS, can be used to test whether the two correlation matrices are invariant, that is, the same. If the test for invariance is nonsignificant, then the two matrices are assumed to be equivalent, with no apparent attrition bias.

Correcting Attrition Bias

Although the strategies used to detect attrition bias are straightforward, there is substantial debate about appropriate strategies to correct attrition bias. Despite the lack of consensus, though, the need for correcting the problem of attrition bias is crucial and continues to motivate statisticians to pursue solutions.

Correction of nonrandom attrition can be broken into two categories. The first category is correction of data when the mechanism of dropping out is known, or in other words, when the researcher knows which characteristics are related to dropping out of the study. The second category is attrition whose causes the researcher does not know.

Known Cause of Attrition

When the cause of attrition is known, the researcher can take steps to control the data analysis procedure to account for the missing data. A model has been developed that simultaneously calculates the research question and the mechanism for missing data. This model is a sample selection model in which two simultaneous regression models are calculated. The first model is a regression model that addresses the research question, with the hypotheses of the study being examined by the regression of the dependent variable on the key independent variables in the study. The second model includes the variables that

are causing attrition, with the dependent variable being a dichotomous variable indicating either continued participation or nonparticipation in the study. The error terms of the substantive dependent variable in the first regression model and the participation dependent variable in the second regression model are correlated. A significant correlation between the two error terms indicates attrition bias. If the correlation is significant, the inclusion of the second model provides corrected regression coefficients for the first, substantive regression model. Thus, the inclusion of the second model that examines attrition bias serves as a correction mechanism for the first, substantive model and enables the calculation of unbiased regression coefficients.

Unknown Cause of Attrition

Heckman proposed a two-step procedure to correct for attrition bias when the cause of the attrition is not readily apparent. He conceptualized the issue of attrition bias as a specification error, in which the variable that accounts for systematic attrition in the study is not included in the regression equation. This specification error results in biased regression coefficients in the analysis. His solution is to first create a proxy of the variable that explains attrition. This is done by conducting a logit regression analysis, similar to the one described in the section on detecting attrition bias. The dependent variable is whether or not each participant participated in the second wave of data collection, and the independent variables are possible variables that may explain or predict dropout. This first step not only tests for attrition bias but also creates an outcome variable, which Heckman calls λ (lambda). Thus, a λ value is computed for all cases in the study, and it represents the proxy variable that explains the causation of attrition in the study.

The second step of Heckman's procedure is to merge the λ value of each participant into the larger data set and then include it in the substantive analysis. In other words, the λ variable is included in the regression equation that is used to test the hypotheses in the study. Including λ in the equation solves the problem

of specification error and leads to more accurate regression coefficients.

While Heckman's model has been used by longitudinal researchers for many years, some concerns have arisen regarding its trustworthiness. Stolzenberg and Relles argue that Heckman's model has been shown to compute inaccurate estimates, and they suggest several cautions when using his model. Nevertheless, Heckman's model offers a possible solution when systematic attrition threatens to bias the results of a study.

—Richard B. Miller and Cody S. Hollist

See also Longitudinal/Repeated Measures Data

Further Reading

- Boys, A., Marsden, J., Stillwell, G., Hatchings, K., Griffiths, P., & Farrell, M. (2003). Minimizing respondent attrition in longitudinal research: Practical implications from a cohort study of adolescent drinking. *Journal of Adolescence*, 26, 363–373.
- Cuddeback, G., Wilson, E., Orme, J. G., & Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30, 19–33.
- Goodman, J. S., & Blum, T. C. (1996). Assessing the nonrandom sampling effects of subject attrition in longitudinal research. *Journal of Management*, 22, 627–652.
- Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow up data. *Journal of Applied Psychology*, 78, 119–128.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Jacobson, J. O. (2004). Place and attrition from substance abuse treatment. *Journal of Drug Issues*, 34, 23–50.
- Mason, M. (1999). A review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement and Evaluation in Counseling and Development*, 32, 111–118.
- Miller, R. B., & Wright, D. (1995). Correction for attrition bias in longitudinal analyses. *Journal of Marriage and the Family*, 57, 921–929.
- Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, 73, 142–167.

AUDIT TRAIL

An audit trail is a collection of documentation that enables tracing the steps of any process or procedure. The term *audit* is commonly applied in accounting situations. In such a context, an audit involves reviewing records of accounting procedures and transactions to assess the validity of financial reports. The assortment of documentation and the pathways available for reviewing that documentation constitute the audit trail. Information technology contexts use audit trails for computer security, to trace, for instance, the path of a system intruder and, ideally, identify the source of any intrusion. Audit trails also are used in food distribution to ascertain the actual nature of food described as organic, for example; to trace the presence or absence of additives; and to locate the origin and pathways of items distributed across various settings. This latter function can be invaluable in public health situations; for example, a food-borne illness can be traced to an infected animal, processing plant, or food service handler.

The audit trail serves a comparable purpose in research applications. In research, the audit trail is used to evaluate decisions and analytic procedures throughout a study to demonstrate the soundness, appropriateness, and in essence the validity of conclusions. One of the original uses of an audit trail in relation to research was in association with large-scale program evaluation projects, and a specific audit process may be included as part of such a study. For example, evaluation of a statewide program to increase graduation rates from high school may involve a review of expenditures, test scores, student progress, graduation rates, and other outcome data. The researchers also might choose to conduct interviews or focus groups with teachers and students involved in the program. To confront questions that the results could be biased by political or personal aims, auditors can be employed to review the procedures of the researchers, ensuring that appropriate data were collected, that conclusions were consistent with the data, and that the results present a valid evaluation of the program under review.

In qualitative studies, because the entire parameters of a study cannot be anticipated in advance, changes often are implemented during a study. The researcher must maintain documentation of decisions and the rationale for any changes as a way to recall and substantiate that such actions were appropriate. In addition to documenting procedures for credibility purposes, an audit trail in qualitative research may include field notes, or notes regarding the behaviors and actions of people and other events happening in the situation where data are collected; methodological documentation; analytic documentation reflecting the researcher's thought processes during data analysis; and documentation of personal responses to capture the investigator's role and reactions as the study progresses. Ongoing developments in software for qualitative data analysis help to consolidate some of these processes by making the creation of field notes and methodological journals a part of the electronic data set.

—Beth Rodgers

See also Authenticity; Text Analysis

Further Reading

- Creswell, J.W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice*, 39(3), 124–130.
- Petterson, M. (2005). The keys to effective IT auditing. *Journal of Corporate Accounting and Finance*, 16(5), 41–46.
- Rodgers, B. L., & Cowles, K. V. (1993). The qualitative research audit trail: A complex collection of documentation. *Research in Nursing and Health*, 16, 219–226.

AUTHENTICITY

Authenticity in education is the fidelity of the intellectual learning environment to the real-world ways in which knowledge is used in the field of study into which the student is being inducted. In other words, the learning that is engendered in formal education and the mechanisms through which such learning is judged are authentic to the extent that there is

congruence between the institutionally derived tasks and reality. In reality, significant accomplishment requires the production, rather than the reproduction, of knowledge. Production of knowledge is a particular type of cognitive work that constructs new knowledge in a meaningful way and has a personal, utilitarian, or aesthetic value beyond the demonstration of competence. Therefore, formal learning tasks are authentic when they meet one or more of criteria for significant accomplishment. Within education, authenticity connotes the quality of intellectual engagement required in reading; writing; speaking; coping with challenges that do not have single solutions; and producing tangible artifacts such as a research report, a musical score, an exhibition of artwork, or a demonstration of an invention. There are three standards by which intellectual engagement can be judged authentic:

1. *Analysis*: The task requires higher-order thinking with content by organizing, synthesizing, interpreting, evaluating, and hypothesizing to produce comparisons, contrasts, arguments, new applications of information, and appraisals of competing perspectives.
2. *Disciplinary concepts*: The task requires an understanding of ideas, concepts, theories, and principles that are central to the academic or professional disciplines into which the student is being inducted.
3. *Elaborated written communication*: The task requires production of detail, qualification, argument, and conclusions that are clear, coherent, and rich.

This interpretation of authenticity is entirely consistent with contemporary theories of learning and knowing, which emphasize how knowledge is represented, organized, and processed in the mind. Because these theories imply that instruction and assessment should be integrated, authenticity can refer to both achievement (the correspondence between classroom instruction and reality) and assessment (the correspondence between instruction and assessment).

Educational practices often assert authenticity through using interactive video environments to engage students in simulated real-life problems. These have produced gains in problem solving,

communication skills, and more positive attitudes to domain knowledge. However, it is not simulations per se but the extent to which they replicate the conditions in which people are challenged in context that determines authenticity. No matter how authentic a task seems, the institutional constraints and policy variables in formal education contexts mean that authenticity in education is an approximation that need not necessarily capture the totality or urgency of all the real-life variables. Currently, the lack of systematic investigation into the effects of the learning context, the learning task, and the learners' interpretations of context and task in the simulated experiences means that the validity and reliability of the simulations are not yet well understood.

—Effie Maclellan

See also Audit Trail

Further Reading

Petraglia, J. (1998). *Reality by design: The rhetoric and technology of authenticity in education*. Mahwah, NJ: Erlbaum.

For information on the Internet, key the term *Authentic Education* into any search engine.

AUTOCORRELATION

Many parametric statistical procedures (e.g., ANOVA, linear regression) assume that the errors of the models used in the analysis are independent of one another (i.e., the errors are not correlated). When this assumption is not met in the context of time-series research designs, the errors are said to be *autocorrelated* or *dependent*. Because time-series designs involve the collection of data from a single participant at many points in time rather than from many participants at one point in time, the assumption of independent errors inherent in many parametric statistical analyses may not be met. When this occurs, the outcome of these analyses and the conclusions drawn from them are likely to be misleading unless some corrective action is taken.

The error in a time-series linear model usually refers to an observed value Y_t (i.e., a dependent variable score observed in a theoretical process at time t) minus the predicted value $\hat{Y}_t$ (based on parameters in the model). When actual sample data are involved (instead of theoretical process data), the predicted values are based on the estimates of the parameters in the model, and the difference $Y_t - \hat{Y}_t$ is called a residual. Hence, a residual is an estimate of an error. For example, if a researcher proposes an ANOVA model for a two-phase interrupted time-series design, the residual is defined as an observed value in a realization (i.e., a sample) of the process minus the mean of the relevant phase. If the sign and size of the residuals are unrelated to the sign and size of the residuals that follow them, there will be no autocorrelation, and this implies that the errors of the model are independent. If, however, positive residuals tend to be followed in time by positive residuals and negative residuals tend to be followed by negative residuals, the autocorrelation will be positive; this is evidence that the independence assumption is violated. Similarly, if positive residuals tend to be followed by negative residuals and negative residuals tend to be followed by positive residuals, the autocorrelation will be negative, and once again, this is evidence that the independence assumption is violated. Autocorrelated errors are especially likely to occur when (a) the time between observations is very short, (b) the outcome behavior changes very slowly, (c) important predictor variables are left out of the model, or (d) the functional form (e.g., linear) of the relationship between the predictors and the outcome is incorrectly specified.

Why Autocorrelation Is Important

Autocorrelation is important because (a) it can affect the validity of inferential statements associated with conventional hypothesis tests and confidence intervals (e.g., positive autocorrelation leads to underestimated p values and confidence intervals that are too narrow), (b) knowledge of its presence can lead a researcher to select a more appropriate statistical analysis, and (c) the precision of predictions made using regression equations can be improved using information regarding autocorrelation.

How Autocorrelation Is Measured

Although one can measure autocorrelation in many different ways, the most frequently encountered method involves the computation of a single coefficient called the *lag-1 autocorrelation coefficient*. This autocorrelation coefficient represents the correlation between the residuals at their associated time t and those same residuals shifted ahead by one unit of time. The sample coefficient computed on actual data is denoted as r_1 whereas the population (or process) parameter is denoted as ρ_1 . Like most two-variable correlation coefficients, the autocorrelation coefficient must fall between -1.0 and $+1.0$. The conventional formula for computing the sample coefficient is

$$r_1 = \frac{\sum_{t=2}^N (e_t)(e_{t-1})}{\sum_{t=1}^N e_t^2},$$

where

e_t is the residual (i.e., the estimate of the error of the model), measured at time t , and

N is the number of residuals in the observed time series.

Consider the data presented in Table 1. The first column lists the time points at which performance measures on a complex job task were obtained. Observations 1 through 10 were obtained during the first (baseline) phase of a two-phase experiment, and observations 11 through 20 were obtained during the second (intervention) phase. The third column contains the residuals, which were computed by subtracting the mean for phase 1 from each observation in phase 1. Similarly, the residuals for phase 2 were computed by subtracting the mean for phase 2 from each observation in phase 2. This approach for computing residuals was used because the investigator chose a simple ANOVA model for the data analysis. This model defines the error as $(Y_t - \mu_t) = \varepsilon_t$ and the estimate of the error (i.e., the residual) as $(Y_t - \bar{Y}_j) = e_t$, where j indicates the phase. The fourth column contains the same residuals shown in column three, except they have been shifted

forward by one time unit. No observation appears in the first row because there is no observation at time point zero. The sum of the values shown in column 5 is 26.3 and the sum of the squared residuals (not shown in the table) is 123.15. The lag-1 autocorrelation coefficient is the ratio of 26.3 over 123.15, which is .21.

Often researchers apply a formal inferential procedure (such as the Durbin-Watson test) to test the hypothesis $\rho_1 = 0$. Using the Durbin-Watson test, we obtain a p value associated with the example autocorrelation coefficient ($r_1 = .21$) that falls above .10, so we have insufficient evidence to conclude that the residuals are autocorrelated. Consequently, the ANOVA model appears acceptable with respect to meeting the independence assumption. Two-phase designs often require more complex models, but regardless of the

Table 1 Values Used in Computing the Lag-1 Autocorrelation Coefficient on the Residuals

Time t	Y_t	e_t	e_{t-1}	$(e_t)(e_{t-1})$
1	5	-1.28	—	—
2	6	-.26	-1.28	.34
3	6	-.25	-.26	.06
4	7	.77	-.25	-.19
5	7	.79	.77	.61
6	8	1.81	.79	1.43
7	4	-2.17	1.81	-3.93
8	7	.85	-2.17	-1.84
9	7	.86	.85	.73
10	5	-1.12	.86	-.97
11	16	1.12	-1.12	-1.25
12	19	4.14	1.12	4.63
13	14	-.85	4.14	-3.50
14	16	1.17	-.85	-.99
15	11	-3.81	1.17	-4.47
16	10	-4.79	-3.81	18.25
17	10	-4.77	-4.79	22.87
18	19	4.25	-4.77	-20.26
19	18	3.26	4.25	13.86
20	15	.28	3.26	.92

Notes: Y_t = observed values (values at $t = 1$ –10 are from Phase 1, and values at $t = 11$ –20 are from Phase 2); e_t = residuals, computed by subtracting the mean for Phase 1 from each observation in Phase 1 and the mean for Phase 2 from each observation in Phase 2; e_{t-1} = residuals shifted forward by one time unit.

complexity of the design or the model, researchers should attempt to determine whether the errors of the model are independent by evaluating the residuals.

More Complete Evaluations of Dependency

The lag-1 autocorrelation coefficient measures the degree of relationship between residuals measured one time period apart. That is, it measures the relationship between residuals of adjacent scores. Although dependency in a time series usually appears in the lag-1 coefficient, this is not always true. Adjacent values may possibly show little or no relationship, while values separated by more than one time unit may show a substantial relationship. One can measure relationships of this type by computing autocorrelation coefficients at lags greater than lag-1. If the autocorrelation is computed between the original series and the original series lagged by two time units, the resulting coefficient is called the lag-2 autocorrelation, denoted r^2 . We can extend this idea to many lags. It is possible to compute K coefficients from N observations, where K is equal to $N - 1$. The whole collection of autocorrelation coefficients (i.e., $r_1, r_2, \dots, r_K$) is called the *autocorrelation function*.

Most time-series computer programs compute autocorrelations for a fraction (usually one sixth to one quarter) of the possible lags. These programs usually represent the coefficients graphically in a correlogram. The vertical dimension of the correlogram indicates the size of the autocorrelation coefficient, and the horizontal dimension indicates the lag. The information displayed in the correlogram is very useful in characterizing the dependency structure in a time series. If all the lagged autocorrelation coefficients in the correlogram hover around zero, this implies that the values in the series are independent. But if large coefficients appear at one or more lags, we have reason to suspect dependency in the series. Most time-series software packages provide formal tests of independence that consider the whole set of coefficients in the correlogram. The most popular of these tests are known as *Box-Pierce* and *Ljung-Box* tests.

Relevant Autocorrelations for Various Time-Series Models

Sometimes investigators estimate autocorrelation on dependent variable scores rather than on errors. It is important to distinguish between autocorrelations estimated on these two types of scores to understand the difference between two popular approaches to time-series analysis. In the approach known as autoregressive, integrated, moving averages (ARIMA) modeling, autocorrelation among *dependent variable scores* is relevant. In the approach known as time-series regression modeling, autocorrelation among *errors of a regression model* is relevant. Although both approaches require the computation of autocorrelation measures, they use information regarding autocorrelation for somewhat different purposes. ARIMA models use autocorrelation measures to identify the type of time-series parameters necessary for modeling the dependent variable scores. In contrast, time-series regression models most often use measures of autocorrelation to determine whether the errors of the regression model show independence. In these regression models, significant autocorrelation should warn researchers that they have misspecified the model (i.e., chosen a substantially wrong model) and that alternative models should be considered. The alternative model may contain parameters that either (a) eliminate autocorrelation among the errors or (b) accommodate that autocorrelation.

—Bradley E. Huitema and Sean Laraway

See also Correlation Coefficient; Time Series Analysis

Further Reading

- Enders, W. (2004). *Applied econometric time series* (2nd ed.). Hoboken, NJ: Wiley.
- Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment*, 10, 253–294.
- Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, 3, 104–116.
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.
- McKnight, S., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze an intervention time

series model with autoregressive error terms. *Psychological Methods*, 5, 87–101.

Software for estimating time series regression models with autoregressive errors (output includes regression coefficients, associated bootstrap tests, and a reduced bias autocorrelation estimate): <http://www.stat.wmich.edu/>, click on Stat Lab → Software → Timeseries. A design matrix and outcome data must be entered.

AVERAGE

To describe a group of values, it is useful to have a typical value or an average value. An average is a summary, so the average value should be representative of a group of values. An average can be used primarily to describe a sample and can also be used to estimate a population value. For example, researchers might want to find an average value in their sample that lets them predict what the average value in the population might be.

There are various measures of central tendency, or average, of a data set, and all have different statistical properties, which makes them sometimes more and sometimes less useful descriptors. Specific problems associated with the distribution the group of values represents are the shape of the distribution (symmetrically distributed or skewed, for example) and the presence or absence of outliers in the data set. Either the average value can be computed, taking all or only some values in the group into account, or it can be a chosen value from the group, seen fit to represent the group.

Most commonly, people refer to the *arithmetic mean* as the average (and vice versa), although this language is ambiguous and should be avoided. The arithmetic mean is the sum of all the values divided by the number of values in the group. There are many variations of the mean, two of the most common ones being the *geometric mean* and the *harmonic mean*. A *trimmed mean* excludes a specific percentage of the upper and the lower end of a distribution, commonly 5% in either direction. A *midmean* is a special case of a trimmed mean in which the mean is calculated for the data between the 25th and 75th percentiles.

The *mode*, the most frequently occurring value, can also be a descriptor of a group of values. The *median* is also a frequently used term for expressing an average value. Especially in skewed data, the mean would be less informative than the median.

It is important to remember that an average value should be presented together with a measure of dispersion.

More about average:

- The mean as the most commonly used average measure is sensitive to extreme scores.
- The median is a suitable average for nonsymmetric distributions and is not affected by outliers.
- The mode is an average value actually represented in the group of values, whereas the mean, as a derivative, can take a value that is not actually represented in the group of values (e.g., the average number of children is 1.2).

—Susanne Hempel

See also Arithmetic Mean; Mean; Measures of Central Tendency; Median; Mode

Further Reading

Molden, D. C., & Dweck, C.-S. (2006). Finding “meaning” in psychology: A lay theories approach to self-regulation, social perception, and social development. *American Psychologist*, 61(3), 192–203.

Descriptive Statistics: <http://www.physics.csbsju.edu/stats/descriptive2.html> (application that computes the mean once data are entered)

Applying Ideas on Statistics and Measurement

The following abstract is adapted from See, W. Y., Wagner, T. H., Shuo, C., & Barnett, P. G. (2003). Average cost of VA rehabilitation, mental health, and long-term hospital stays. *Medical Care Research and Review*, 60(3 suppl), 40S–53S.

One of the most common methods used to better understand a collection of data points is through the calculation of an **average** (which includes such descriptive statistics as the mean, the mode, and the median). In this article, researchers Wei You See and his colleagues at

Stanford University describe the development of a database for the cost of inpatient rehabilitation, mental health, and long-term care stays in the Department of Veterans Affairs from fiscal year 1998. As a unit of analysis, they used bedsection, which is similar to a hospital ward, and they classified inpatient services into nine categories, including rehabilitation, blind rehabilitation, spinal cord injury, psychiatry, substance abuse, intermediate medicine, domiciliary, psychosocial residential rehabilitation, and nursing home. For each of these nine categories, they estimated a national and a local average per diem cost. The next step was to calculate what they call encounter-level costs, which was done by multiplying the average per diem cost by the number of days of stay in the fiscal year. Their conclusion? The national cost estimates for hospitalization are more reliable than the local cost estimates for the same.

AVERAGE DEVIATION

The *average deviation* (AD) is used as a measure of dispersion or within-group interrater agreement and may be referred to as the *average absolute deviation* or *mean deviation*. The average deviation is often defined in one of two ways: by deviations from the mean (AD_M) or by deviations from the median (AD_{Md}). The average deviation is calculated by taking the difference between each score and the mean (or median), summing the absolute values of these deviations, and then dividing the sum by the number of deviations. As a measure of dispersion, the larger the AD, the greater is the variability in a distribution of scores. As a measure of within-group interrater agreement, the larger the AD, the greater is the disagreement among raters evaluating a single target on a categorical rating scale.

The formula for the computation of the average deviation using the mean is

$$AD_M = \frac{\sum |X - \bar{X}|}{n}$$

where

Σ directs you to add together what follows it,

X is each individual score in the distribution of scores,

$\bar{X}$ is the mean,

the vertical lines are the absolute value symbols and direct you to disregard the fact that some deviations are positive and some negative, and

n is the number of cases or number of raters.

The formula for the computation of average deviation using the median (AD_{Md}) would substitute the median for the mean in the above equation.

More about the average deviation as a measure of dispersion:

- It gives equal weight to the deviation of every value from the mean or median.
- The average deviation from the median has the property of being the point at which the sum of the absolute deviations is minimal compared with any other point in the distribution of scores.
- Given that the AD is based on every value in the distribution of scores, it provides a better description of the dispersion than does the range or quartile deviation.
- In comparison with the standard deviation, the AD is less affected by extreme values and easier to understand.

More about the average deviation as a measure of within-group interrater agreement:

- The AD provides an index of interrater agreement in the metric (measurement units) of the original rating scale.
- A statistically derived cutoff for an acceptable level of disagreement in raters' evaluations of a single target is $c/6$ values, where c is the number of response options or rating categories. Values of AD exceeding this cutoff value (e.g., values of AD exceeding a cutoff value of 1.2 on a 7-point rating scale) would indicate disagreement among raters, and values of AD below the cutoff would indicate agreement in raters' scores of the single target.
- In comparison with other measures of within-group interrater agreement, including the standard

deviation, the AD index is easiest to understand and interpret.

—*Michael J. Burke*

See also Standard Deviation; Variance

Further Reading

Dunlap, W. P., Burke, M. J., & Smith Crowe, K. (2003). Accurate tests of statistical significance for $r_{\text{sub(WG)}}$ and average deviation interrater agreement indexes. *Journal of Applied Psychology*, 88(2), 356–362.

Averages and deviations: http://www.sciencebyjones.com/average_deviation.htm

B

Although he may not always recognize his bondage, modern man lives under a tyranny of numbers.

—Nicholas Eberstadt

BABBAGE, CHARLES **(1791–1871)**

Charles Babbage is best known as the Father of Computing, having formulated the idea of a mechanical calculator during his student days.

Babbage was born in London on Boxing Day, December 26, 1791. His father was a banker from Devon, and Babbage was educated in Totnes, with the family moving to Teignmouth in 1808. Babbage was a sickly child who therefore received much of his education from private tutors at home. One consequence was that, on arriving at Cambridge University in 1810, he was somewhat disappointed with the standards expected. In 1812, with fellow undergraduates, he established the Analytical Society with the aim of promoting the use of our modern notation for differential calculus over the Newtonian presentation then in vogue.

He graduated in 1814, the year of his marriage to Georgiana Whitmore (who died in 1827 following the birth of their eighth child—only three survived to adulthood). In 1816, following the publication of two major papers, and as a consequence of his work with the Analytical Society, Babbage was elected a Fellow

of the Royal Society. In 1820, he was elected a Fellow of the Royal Society of Edinburgh and was a founding member of what is now the Royal Astronomical Society. In 1828, he was appointed Lucasian Professor of Mathematics at Cambridge University (a post he held until 1839). In 1834, the decision to found a statistics society (now the Royal Statistical Society) was made at his house.

Babbage conceived the notion of constructing a machine for automated calculations while at Cambridge, but it was not until 1822 that he completed a working model. This machine was able to calculate the values of $n^2 + n + 1$ for successive values of n at the rate of 12 a minute. The machine worked by using differences: thus the successive values of $n^2 + n + 1$ are 3, 7, 13, 17, . . . , with constant differences of 4. The machine that Babbage conceived is therefore known as a *difference engine*. The initial model was well received, and Babbage received a grant to build a bigger, more powerful engine. However, as is so often the case with construction projects, the costs escalated enormously, and the project was finally abandoned in 1834.

Babbage then turned his attention to the construction of a more versatile *analytical engine*, which used punched cards adapted from a Jacquard loom (which

enabled intricate woven patterns in cloth). The initial, 1835 design was for a machine five meters (16.4 feet) tall. For the next 10 years, the plans were refined, and it was at this time that Babbage corresponded with Ada, Countess of Lovelace, in the construction of algorithms for the embryonic machine—in other words, the world’s first computer programs. Babbage died in London on October 18, 1871.

—Graham Upton

See also Probability Sampling

Further Reading

Swade, D. (2002). *The difference engine: Charles Babbage and the quest to build the first computer*. New York: Penguin.

Babbage’s machines: <http://www.sciencemuseum.org.uk/on-line/babbage/index.asp>

BAR CHART

A bar chart is a specific type of chart that visually represents data as a series of horizontal bars, with the Y axis representing the categories contained in the data and the X axis representing the frequency. It is different from a column chart in that column charts display the data vertically.

Bar charts are most often used for categorical data that is, by definition, not dynamic in nature. For example, if one were interested in an examination of sales figures by brand before and after a marketing campaign, a bar chart would be the appropriate way to illustrate such information, as shown in the following example. First, here are the data.

	Brand X	Brand Y
Before	56.3	76.8
After	97.4	87.5

Note: Figures represent sales in millions of dollars.

A bar chart is relatively simple to construct manually. Following these steps and using graph paper is the easiest way to be accurate.

1. Group the data as shown in the above example.
2. Define the Y axis as “Brand.”
3. Indicate on the X axis the scale that is to be used, which in this case is millions of dollars, ranging from 55 to 100.
4. Draw each bar for each brand to correspond with the data, making sure that the bars are appropriately colored or patterned so they are easily distinguishable from one another.

Using Excel, the process is much simpler.

1. Create the data on a new worksheet and save it.
2. Using the mouse, select all the data, including the column and row headings.
3. Click on the Chart Wizard icon on the Excel toolbar.
4. Click Finish in the dialog box.

While the finished bar chart may not appear as attractive as you might like, modifications are relatively easy to make, as shown in Figure 1.

The following changes were made:

- Major gridlines were removed.
- All coloring was deleted except for the gray added to one of each pair of bars to help distinguish it from the other.
- Value labels were added at the end of each bar.
- A title was added, as were labels for each axis.

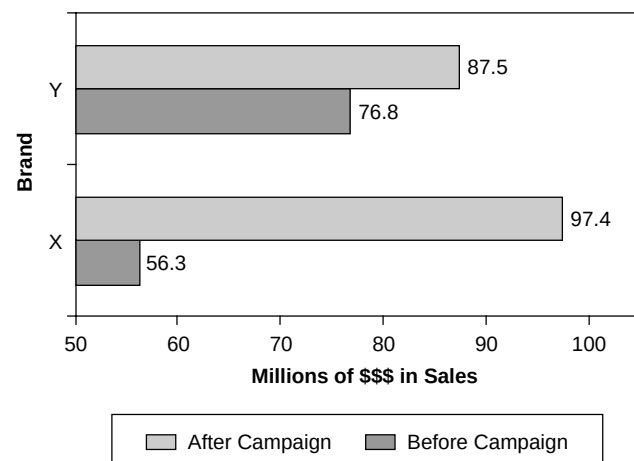


Figure 1 Sales by Brand

- The X axis minimum and maximum were changed from Excel's default values to 50 and 105.
- The border was removed from the legend.

—Neil J. Salkind

See also Area Chart; Line Chart; Pie Chart

Further Reading

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

BASAL AGE

Basal age represents the highest level, on a test standardized in units corresponding to mental age or age-equivalents, below which it can be assumed that all items would be answered correctly. For example, a child who correctly answers the required number of items at a certain age level would be given credit for all preceding items on the test even though the child has not actually been tested on those items. The term *basal age* is often used interchangeably with *basal score* or *basal*.

The point at which to begin the calculation of basal age is usually estimated from a child's chronological age or, for children with learning or language disabilities, from a functional estimate of age. For children who demonstrate considerable scatter in their performance or who perform at significantly different levels from their age-matched peers, the calculation of a basal age makes it easier to determine a meaningful starting point on a test.

Although tests vary somewhat in how basal scores are calculated, the procedure is usually very similar. The Stanford-Binet Intelligence Scales (fifth edition), for example, uses basal levels in all its 11 testing blocks. Testing is begun at a block that is believed to represent a child's general ability level. If a basal is not obtained, testing moves back to successively lower blocks until a basal is established. On the Peabody Picture Vocabulary Test (third edition), a basal is established when a respondent makes no or only one error within a set of items. On the Woodcock Language Proficiency Battery-Revised, the number of

correct responses required to establish a basal age varies between four and six, depending on the particular subtest.

On almost all tests, the examiner administers test items in reverse order from the estimated starting point until a basal is obtained in the manner specified. In the event that no basal level is achieved, the first item on the test is usually considered to represent the basal age.

Tests that employ basal scores typically also use *ceiling scores*, or previously defined accuracy levels that determine the point at which a test is terminated. Thus, the basal and ceiling essentially define an examinee's functional range. Moreover, they serve to limit the number of test items that are administered. Usually, age-equivalent scores are calculated based on the raw score derived from the ceiling item and the number of correct responses below it.

—Carole E. Gelfer

Further Reading

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

BASIC PERSONALITY INVENTORY

The Basic Personality Inventory (published by Sigma Assessment Systems, www.sigmaassessmentsystems.com) was developed by Douglas N. Jackson and is currently in its second edition (BPI-II). The Basic Personality Inventory was derived through factor analysis of the 28 scales of the Differential Personality Inventory, and the resulting 11 dimensions identified were augmented with the critical item scale deviation to form the basis for a multiscale inventory of psychopathology and psychosocial adjustment. Once the 12 dimensions to be measured had been defined as constructs, new items were written and selected for the scales. The result was 20 items per scale, balanced in terms of true and false keying for 11 of the 12 scales (all 20 deviation items are true keyed). There is no item overlap between scales. Items selected correlated more highly with their own scale than with any

other scale. In order to foster local independence, the items are ordered such that one item is presented from each of the 12 scales in succession, and then this block of 12 items repeats in the same sequence with alternative keying.

The 12 scales are grouped according to three primary factors: internalizing psychopathology (hypochondriasis, anxiety, thinking disorder, persecutory ideas, and deviation), affective psychopathology (depression, social introversion, and self-depreciation), and externalizing psychopathology (deviation, interpersonal problems, alienation, and impulse expression). The denial scale is not included in these factors because it is used for validity purposes as a measure of the dimension representing the Differential Personality Inventory scales of repression, shallow affect, and defensiveness. Deviation is also used as a critical item index of deviant behaviors requiring further exploration, such as suicidal ideation or substance abuse. In addition to the 12 standard scales, a 20-item supplementary social desirability scale was formed by taking the most desirable item and most undesirable item from 10 of the 12 scales (validity scales denial and deviation excluded). Additional indices that can be scored for validity assessment are number of true responses, perseveration, person reliability, fake good, and fake bad.

The BPI-II is available in English, French, and Spanish and can be administered in paper-and-pencil format with a reusable question booklet and answer sheet to individuals or to groups in about 35 minutes. Computerized administration with automated scoring and reporting is also available. Reports generated by the software are either clinical with some limited interpretation or ASCII (text) data file reports for research. Mail-in and Internet scoring and report generation are available. The reading level is estimated to be between Grades 5 and 6. Normative data stratified by gender are available separately for juveniles and adults.

—John R. Reddon

See also Comrey Personality Scales; Jackson, Douglas N.; Minnesota Multiphasic Personality Inventory; NEO Personality Inventory; Personality Tests

Further Reading

- Holden, R. R., & Jackson, D. N. (1992). Assessing psychopathology using the Basic Personality Inventory: Rationale and applications. In J. Rosen & P. McReynolds (Eds.), *Advances in psychological assessment* (Vol. 8, pp. 165–199). New York: Plenum.
- Kroner, D. G., Black, E. R., & Reddon, J. R. (1996). Item content of the Basic Personality Inventory. *Multivariate Experimental Clinical Research*, 11, 61–73.

Basic Personality Inventory: <http://www.sigmaassessmentsystems.com/assessments/bpi.asp>

BASIC RESEARCH

Basic research consists of empirical studies that “answer fundamental questions about the nature of behavior” (Cozby, 1985, p. 8). Its main goal is the understanding of a phenomenon. Basic research is less concerned with understanding and solving practical problems, which are the primary foci of applied research. Instead, basic research strives to develop a body of knowledge that has no obvious or immediate practical application. However, this knowledge can lead to interventions that alleviate human problems and distress. For example, Skinner’s research with pigeons in the 1930s eventually led to the development of theoretical principles that had a profound impact on clinical and educational practices.

Basic research is most effective and influential when its explicit goal is the development of theory. Although research that is conducted solely for the sake of knowledge represents valid and worthy basic research, research that tests hypotheses that are deductively derived from a theory offers much greater promise of advancing science. Indeed, the primary purpose of basic research is the testing of hypotheses that are generated by a particular theory. As theory-based hypotheses are rejected and supported, theories are validated, rejected, and refined.

The interaction among research, theory, and application is represented in Figure 1, which is a model adopted from Olson. Theories are developed to explain human behavior and interaction. These

theories are then applied to practical situations, such as childhood behavioral problems or distressed marriages. For example, the theoretical principles of childhood attachment have been extended to adult attachment, and Emotionally Focused Therapy has been developed from these principles to assess and treat conflicted romantic relationships. Within this therapy model, relationship difficulties are conceptualized from an attachment perspective, and treatments have been developed that seek to repair attachment injuries and promote secure attachment in relationships.

As illustrated in Figure 1, the purpose of basic research is to test the validity of particular theories. In the case of attachment theory, a large body of empirical research has accumulated that provides robust evidence for the general principles of the theory. This empirical support has validated attachment theory, providing important credibility to the techniques that were derived from the theory to assess and treat couples.

In some cases, research has rejected significant theoretical propositions. For example, for many years the dominant theoretical conceptualization of autism was that it was caused by poor parenting. Based on this theory of autism, treatment focused on providing adequate care outside the realm of the family. Subsequent research has consistently failed to support this theory, and more recent research has supported theories that link autistic symptoms to neurological impairment. With the conceptualization of autism as a neurological disorder, parents are included as essential partners in the treatment of their child because

they provide consistency between behavioral treatments at school and at home.

Thus, basic research serves as a scientific watchdog by providing vital empirical evidence in the process of developing and validating theories. Without this validation role, science stagnates because theory development is curtailed. The system of checks and balances between theoretical generation and basic research enables the continual refinement of valid theories. These theories can then be deductively applied to a variety of human and social problems.

Figure 1 also illustrates the role of applied research, which has a primary aim of directly testing the effectiveness of applications and interventions. Multiple clinical outcome studies have demonstrated that Emotionally Focused Therapy, which is based on attachment theory, is effective in improving relationship functioning among couples. Thus, basic and applied research are complementary, with basic research examining the validity of theories and applied research testing the effectiveness of applications and interventions that are derived from validated theories.

—Richard B. Miller and Ryan G. Henry

See also Applied Research

Further Reading

- Bettelheim, B. (1967). *The empty fortress*. New York: Free Press.
- Cozby, P. C. (1985). *Methods in behavioral research* (3rd ed.). Palo Alto, CA: Mayfield.
- Evans, J. D. (1985). *Invitation to psychological research*. New York: Holt, Rinehart, & Winston.
- Handleman, J. S., & Harris, S. L. (1986). *Educating the developmentally disabled: Meeting the needs of children and families*. San Diego, CA: College-Hill Press.
- Johnson, S. M. (1996). *The practice of emotionally focused marital therapy: Creating connections*. New York: Brunner/Mazel.
- Johnson, S. M., Hunsley, J., Greenberg, L., & Schindler, D. (1999). Emotionally focused couples therapy: Status and challenges. *Clinical Psychology: Science & Practice*, 6, 67–79.
- Olson, D. H. (1976). *Treating relationships*. Lake Mills, IA: Graphinc.

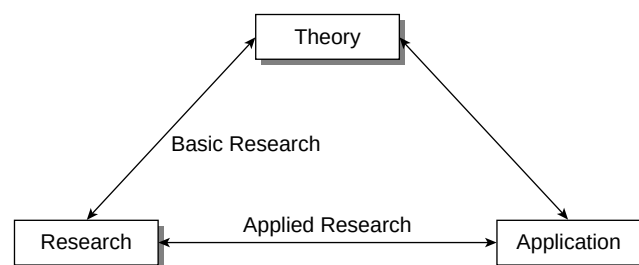


Figure 1 The Relationship Between Theory, Research, and Application

Source: Adapted from Olson, 1976 (p. 565).

Shaughnessy, J. J., & Zechmeister, E. B. (1994). *Research methods in psychology* (3rd ed.). New York: McGraw-Hill.
 Spata, A. V. (2003). *Research methods: Science and diversity*. New York: Wiley.

BAYES FACTORS

The problem of comparing multiple competing models is a difficult one and has been the focus of much research over the years. For statisticians of a Bayesian inclination, the *Bayes factor* offers an easy way of comparing two models. Simply defined, the Bayes factor is the ratio of the posterior and the prior odds. More specifically, denote the data as $x = x_1, x_2, \dots, x_n$ and the two models as M_1 and M_2 , respectively. The Bayes factor in favor of model M_1 , and thus against model M_2 , is

$$B_{12}(x) = \frac{p(M_1|x)}{p(M_2|x)} / \frac{p(M_1)}{p(M_2)},$$

where $p(M_i)$ is the prior probability assigned to model i and $p(M_i | x)$ is the posterior probability of model i after observing the data.

Since the posterior probability of model i can be expressed as $p(M_i | x) = p(x | M_i) p(M_i)$, the Bayes factor is also sometimes given as $B_{12}(x) = p(x | M_1) / p(x | M_2)$. It follows that the posterior odds are equal to the Bayes factor times the prior odds.

The first representation provides an intuitive explanation of what the Bayes factor measures, namely, how the data have affected the odds in favor of model M_1 . If $B_{12}(x) > 1$, then the posterior odds in favor of the first model are higher than the prior odds in favor of the first model. In other words, the data have increased our relative amount of belief in the first model. If, on the other hand, $B_{12}(x) < 1$, the posterior odds in favor of model M_1 have decreased on observing the data.

In practice, values of $B_{12}(x)$ smaller than 3 are often taken as providing no evidence in favor of M_1 over M_2 , values of $B_{12}(x)$ between 3 and 20 give positive evidence, values between 20 and 150 are indicative of strong evidence, and any value of the Bayes factor over 150 is taken to be very strong evidence. Although

these values are only guidelines, they are useful for the calibration of results.

One advantage of Bayes factors over traditional approaches to model comparison is that the models do not have to be nested, as this simple example demonstrates. Suppose we have data $x_1, \dots, x_n$, independent, identically distributed, coming from either a negative binomial distribution with probability p of success (this is model M_1) or from a Poisson distribution with mean λ (this is model M_2). We use different notation for the parameters of these two models to emphasize that there is no need for the models under consideration to be related to each other in any way. Both these models are completely specified, and so the Bayes factor for comparing the two models is simply the likelihood ratio, that is

$$B_{12}(x) = \frac{p^n(1-p)^{n\bar{x}}}{\lambda^{n\bar{x}} e^{-n\lambda} (\prod x_i!)^{-1}}.$$

Now, suppose that p and λ are not known. To carry out a Bayesian analysis, it is necessary to assign prior distributions to the unknown parameters of the two models. For simplicity, suppose that $p \sim \text{Beta}(\alpha_1, \beta_1)$ and $\lambda \sim \text{Gamma}(\alpha_2, \beta_2)$. Then it can be shown that

$$p(x|M_1) = \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n + \alpha_1)\Gamma(n\bar{x} + \beta_1)}{\Gamma(n + n\bar{x} + \alpha_1 + \beta_1)}$$

and

$$p(x|M_2) = \frac{\Gamma(n\bar{x} + \alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n + \beta_2)^{n\bar{x} + \alpha_2}} \frac{1}{\prod x_i!};$$

the Bayes factor is the ratio of these two. In this case, we need to have information on α_1 , β_1 , α_2 , and β_2 in order to evaluate the Bayes factor, which will always depend on the prior specification. As we assign different values to these four parameters, resulting in different prior distributions (representing our different opinions about p and λ and how sure we are of those opinions), the Bayes factor will in turn vary.

A simple numerical example demonstrates this. Table 1 shows values of the Bayes factor for three different sample configurations and three choices of

Table 1 Values of the Bayes Factor for Comparing the Negative Binomial and Poisson Models, With Different Data and Prior Configurations

	$\alpha_1 = 3, \beta_1 = 6$ $\alpha_2 = 6, \beta_2 = 3$	$\alpha_1 = 30, \beta_1 = 60$ $\alpha_2 = 60, \beta_2 = 30$	$\alpha_1 = 1, \beta_1 = 4$ $\alpha_2 = 8, \beta_2 = 2$
$\sum_{i=1}^4 x_i = 0$	4.89	25.59	93.73
$\sum_{i=1}^4 x_i = 8$	0.09	0.09	0.12
$\sum_{i=1}^4 x_i = 16$	0.08	0.22	0.02

parameters for the prior distributions. For ease of calculation, the size of the sample is only 4, and in each of the three cases, $x_1 = x_2 = x_3 = x_4$. The priors are chosen so that $\alpha_1 \alpha_2 = \beta_1 \beta_2$. As can be seen, the Bayes factor can change quite dramatically as the nature of the data or of the prior is changed.

The effect of the sample is seen in particular by comparing the first and second lines of the table. The Bayes factor is dramatically reduced, no matter what the choice of priors. Indeed, looking at the second and third lines, the practical effect of the prior is minimal—the same conclusion would be reached in all these cases, that the data do not speak in favor of the negative binomial model. The top line of the table tells a very different story. There, the negative binomial model is preferred to the Poisson model, with moderately positive to strongly positive evidence in favor of the former. The value of the Bayes factor and the conclusions we would draw from the analysis depend quite heavily for this sample configuration on the prior distribution. Even though for all three priors the same model would be indicated, the strength of our belief in that model would vary greatly.

In spite of the convenience of the Bayes factors for comparing the strength of evidence in favor of different, possibly non-nested, models, they are not without drawbacks. Among the major criticisms leveled at the Bayes factor is that it never loses its dependence on the prior specification, in contrast to Bayesian inferential procedures, in which the influence of the prior distribution weakens as more data are collected. Another critique of the Bayes factor is that it corresponds to a zero-one loss on the decision “Which of these two models pertains?” What this means is that if the wrong choice is made, that is, the wrong model is chosen, it

doesn’t matter how far off the choice is. This does not correspond to the way in which statisticians usually think about model selection problems. Furthermore, Bayes factors are hard to calculate and to interpret if improper priors are used.

Because of these various difficulties with the ordinary Bayes factor, researchers have developed a range of alternatives, most of which aim at some

form of automatic model selection, particularly to avoid the problems associated with the dependence on priors. These include intrinsic Bayes factors (arithmetic and geometric), fractional Bayes factors, posterior Bayes factors, and others. However, these have also been criticized on the grounds of being arbitrary, of lacking any real Bayesian justification, and of avoiding the difficult issue of prior choice altogether.

The *Bayesian information criterion* (BIC) is related to the Bayes factor and is useful for model comparison in its own right. The BIC of a model is defined as $-2 (\log \text{maximized likelihood}) + (\log n)(\text{number of parameters})$. BIC penalizes more-complex models (those with many parameters) relative to simpler models. This definition permits multiple models to be compared at once; the model with the highest posterior probability is the one that minimizes BIC. The BIC can also be derived by an approximation to the logarithm of the Bayes factor, given by the *Schwarz criterion* for comparing two models,

$$S = \log p(x|\hat{\theta}_1, M_1) - \log p(x|\hat{\theta}_2, M_2) - \frac{1}{2}(d_1 - d_2) \log n,$$

where

$\hat{\theta}_i$ is the maximum likelihood estimator for the parameter q_i under model M_i ,

d_i is the dimension of θ_i , and

n is the sample size.

One computational advantage of this approximation is that there is no need to introduce a prior into the

calculation at all. However, it is only a rough approximation, and if a detailed analysis is required, it will generally not be suitable. The Schwarz criterion multiplied by -2 is the BIC for the comparison of two models.

In summary, the Bayes factor inherits the strengths of the Bayesian paradigm, namely, a logical foundation that transfers easily to new situations. In addition, Bayes factors allow researchers to compare nonnested models and to incorporate prior information or beliefs about a theory into the testing situation. On the other hand, Bayes factors are heavily dependent on the prior specifications, even for large sample sizes, correspond to an “all or nothing” approach to model comparison, and can be difficult to calculate.

—Nicole Lazar

See also Bayesian Information Criterion; Bayesian Statistics

Further Reading

- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, UK: Wiley.
- Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90, 773–795.

BAYESIAN INFORMATION CRITERION

The Bayesian information criterion (BIC) is a statistic used for comparison and selection of statistical models. BIC is given by a simple formula that uses only elements of standard output for fitted models. It is calculated for each model under consideration, and models with small values of BIC are then preferred for selection. The BIC formula and the sense in which the model with the smallest BIC is the “best” one are motivated by one approach to model selection in Bayesian statistical inference.

Definition

Suppose that we are analyzing a set of data D of size n . Here n is the sample size if D consists of statistically independent observations and the “effective sample size” in some appropriate sense when the observations are not independent. Suppose that alternative models M_k are considered for D , and that each model is fully specified by a parameter vector θ_k with p_k parameters. Let $p(D|\theta_k; M_k)$ denote the likelihood function for model M_k , $l(\theta_k) = \log p(D|\theta_k; M_k)$ the corresponding log-likelihood, and $\hat{\theta}_k$ the maximum likelihood estimate of θ_k .

Let M_s denote a saturated model that fits the data exactly. One form of the BIC statistic for a model M_k is

$$\begin{aligned} \text{BIC}_k &= -2[l(\hat{\theta}_k) - l(\hat{\theta}_s)] - df_k \log n \\ &= G_k^2 - df_k \log n, \end{aligned} \quad (1)$$

where

$l(\hat{\theta}_s)$ is the log-likelihood for the saturated model,

G_k^2 is the deviance statistic for model M_k , and

df_k is its degrees of freedom.

This version of BIC is most appropriate when the idea of a saturated model is natural, such as for models for contingency tables and structural equation models for covariance structures. The deviance and its degrees of freedom are then typically included in standard output for the fitted model. In other cases, other forms of BIC may be more convenient. These variants, all of which are equivalent for purposes of model comparison, are described at the end of this entry.

Motivation as an Approximate Bayes Factor

The theoretical motivation of BIC is based on the idea of a Bayes factor, which is a statistic used for comparison of models in Bayesian statistical analysis. First, define for model M_k the integrated likelihood

$$p(D|M_k) = \int p(D|\theta_k, M_k) p(\theta_k|M_k) d\theta_k, \quad (2)$$

where $p(\theta_k | M_k)$ is the density function of a prior distribution specified for the parameters θ_k , and the integral is over the range of possible values for θ_k . Defining $p(\theta_k | M_s)$ similarly for the saturated model, the Bayes factor between models M_s and M_k is the ratio $BF_k = p(D | M_s) / p(D | M_k)$. It is a measure of the evidence provided by the data in favor of M_s over M_k . The evidence favors M_s if BF_k is greater than 1 and M_k if BF_k is less than 1.

BIC_k is an approximation of $2\log BF_k$. The approximation is particularly accurate when each of the prior distributions $p(\theta_k | M_k)$ and $p(\theta_s | M_s)$ is a multivariate normal distribution with a variance matrix comparable to that of the sampling distribution of the maximum likelihood estimate of the parameters based on a hypothetical sample of size $n = 1$. An assumption of such prior distributions, which are known as *unit information priors*, thus implicitly underlies BIC Equation 1. Their motivation and the derivation of BIC are discussed in detail in the Further Reading list below.

A positive value of BIC_k indicates that the saturated model M_s is preferred to M_k (i.e., that $BF_k > 1$), and a negative BIC_k indicates that M_k is preferred. Values of BIC can also be compared between different nonsaturated models, and the model with the smaller BIC is then preferred. The model with the smallest value of BIC overall is regarded as the best model in the Bayes factor sense of being supported most strongly given the data D and the prior distributions of the parameters of all the models.

Example

For an illustration of the use of BIC, consider the data in Table 1. This shows the cross-tabulation of the passengers of RMS *Titanic*, classified according to sex (men vs. women or children), the class they traveled in (first, second, or third), and whether they survived

Table 1 Passengers of the *Titanic*, Classified According to Passenger Class, Sex, and Survival Status

Class	Group	Survivor		Total
		Yes	No	
First	Man	57 (0.326)	118 (0.674)	175 (1.000)
	Woman or child	146 (0.973)	4 (0.027)	150 (1.000)
	Total	203 (0.625)	122 (0.375)	325 (1.000)
Second	Man	14 (0.083)	154 (0.917)	168 (1.000)
	Woman or child	104 (0.889)	13 (0.111)	117 (1.000)
	Total	118 (0.414)	167 (0.586)	285 (1.000)
Third	Man	75 (0.162)	387 (0.838)	462 (1.000)
	Woman or child	103 (0.422)	141 (0.578)	244 (1.000)
	Total	178 (0.252)	528 (0.748)	706 (1.000)
Total		499 (0.379)	817 (0.621)	1316 (1.000)

Note: Numbers in parentheses are proportions within the rows.

the sinking of the ship. Table 2 shows results for some models fitted to these data. These are standard loglinear models, identified in the second column of Table 2 using conventional concise notation for such models. For example, the expression (CY, SY) for model 5 indicates that it includes the two-way interactions between class (C) and survival (Y) and between sex (S) and survival, but no interaction between class and sex. The saturated model here is model 10.

The model with the smallest BIC is model 9, for which $BIC = -7.1$. This is also the only model with a negative BIC, i.e., the only one preferred to the saturated model. Model 9 includes all one-way and

Table 2 Results for Loglinear Models Fitted to the Data in Table 1

Model	Terms	G^2	df	BIC
1	(C, S, Y)	582.1	7	531.8
2	(C, SY)	229.7	6	186.6
3	(S, CY)	449.4	5	413.5
4	(CS, Y)	568.8	5	532.9
5	(CY, SY)	97.0	4	68.3
6	(CS, SY)	216.4	4	187.7
7	(CS, CY)	436.2	3	414.7
8	(CS, CY, SY)	89.1	2	74.7
9	8 + (Class 3)*S*Y	0.051	1	-7.1
10	(CSY)	0	0	0

Notes: C = class; S = sex; Y = survival status; G^2 = goodness of fit; df = degree of freedom; BIC = Bayesian information criterion.

two-way interactions, and the three-way interaction between sex, survival, and traveling in third class. In other words, the model specifies two patterns of association between sex and survival, one for third-class passengers and one jointly for first- and second-class passengers. Considering Table 1, it appears that the main difference between these groups was that there was a smaller disparity between men's chances of survival and those for women and children in the third class than there was in the other two classes.

Relations to Other Methods of Model Selection

In BIC Equation 1, the deviance G_k^2 is a measure of the goodness of fit of a model, with well-fitting models having small values of G_k^2 . In the second term, df_k is a decreasing function of the number of parameters p_k , which can be regarded as a measure of the complexity of the model. The term $-df_k \log n$ is known as the penalty term of BIC because it is an increasing function of p_k and thus "penalizes" a model for its complexity. Considering increasingly complex models will generally lead to a decrease in the deviance but an increase in the penalty term. The terms thus pull in different directions, apparently expressing a trade-off between fit and complexity. Small values of BIC are

obtained for models that achieve a good balance between these two, or in other words, a good fit with relatively little complexity.

Other choices of the penalty term give different penalized model selection criteria of the same general form. The most common of these is the Akaike information criterion, where the $-df_k \log n$ in Equation 1 is replaced by $-2df_k$. The Akaike information criterion is used for model selection in broadly the same way as BIC even though the two statistics have very different theoretical motivations.

BIC and other penalized criteria are often used as complements to standard significance tests in model selection. For example, for the models in Table 2, both the deviances G_k^2 and their differences between nested models can be used as test statistics for this purpose. In this example, both BIC and significance tests identify model 9 as the preferred model, but in general they will not always be in agreement. In particular, conclusions often differ in large samples, where significance tests are sensitive to even small observed discrepancies and may reject most models as having significant lack of fit. The penalty term of BIC offsets some of this tendency, so BIC will often favor less-complex models more than goodness-of-fit tests do.

Alternative Formulas for BIC

Forms of BIC other than Equation 1 may be more convenient when the deviance G_k^2 is not immediately available. For many regression models, it is easier to replace the saturated model as the baseline for comparisons with a null model M_0 , which includes no explanatory variables. This gives the statistic

$$\begin{aligned} BIC'_k &= -2[l(\hat{\theta}_k) - l(\hat{\theta}_0)] + df'_k \log n \\ &= -LR_k + df'_k \log n, \end{aligned} \quad (3)$$

where

$l(\hat{\theta}_0)$ is the log-likelihood for the null model,

LR_k is the likelihood ratio test statistic for testing M_0 against M_k , and

df'_k is its degrees of freedom.

For example, for linear regression models, $BIC'_k = n \log(1 - R_k^2) + p'_k \log n$, where R_k^2 is the standard R^2 statistic for model M_k and p'_k denotes the number of explanatory variables (excluding the intercept term) in M_k . In terms of the Bayesian motivation, BIC'_k is an approximation of $2 \log[p(D | M_0)/p(D | M_k)]$.

When standard computer output includes only the log-likelihood $l(\hat{\theta}_k)$ instead of G_k^2 or LR_k , the most convenient BIC-type statistic is simply $-2l(\hat{\theta}_k) + p_k \log n$. This is an approximation of $2 \log p(D | M_k)$ under the unit information prior for θ_k discussed above. Models with small values of this statistic are again preferred, as they are for Equations 1 and 3. All these three variants of BIC will always lead to the same conclusions about preferences among models and the selection of the best model.

—Jouni Kuha

See also Akaike Information Criterion; Bayes Factors

Further Reading

- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 188–229.
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–163). Cambridge, MA: Blackwell.
- Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27, 359–397.

BAYESIAN STATISTICS

The term *Bayesian statistics* refers to the field of statistical analysis that deals with the estimation of probability distributions for unobserved or “latent” variables based on observed data. When a researcher

collects data from an educational assessment, for example, the test score itself is not typically of interest, but rather the “trait” or “ability” that is thought to underlie and influence an examinee’s responses to the test questions. Indeed, most measurements in the social sciences are collected as substitutes for some latent variable or variables that cannot be observed directly. That is, the data (denoted x) are not of as much interest to the researcher as the true parameter values (denoted θ) that gave rise to the data. Under the framework of a Bayesian data analysis, statistical inferences are therefore based on a quantity that is of direct interest to the analyst (i.e., θ), not some proxy for that quantity of interest (i.e., the data, x).

Bayesian statistical analysis is named after its founder, Thomas Bayes (1702–1761), an 18th-century minister and mathematician who first introduced what is now known as Bayes’ theorem or Bayes’ rule. Bayes’ theorem posits that the conditional probability of an event, A, occurring, given that another event, B, has occurred, is a function of the joint probability of A and B (the probability of events A and B co-occurring) divided by the marginal probability of B. Given this result, the conditional probability of A given B can be stated as the conditional probability of B given A, multiplied by the marginal probability of A, divided by the marginal probability of B:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

This probability statement can be used in a variety of situations for determining the conditional probability of a given event. Applications of Bayes’ theorem are especially useful to statisticians when the theorem is phrased in terms of distributions of observed and unobserved variables:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}.$$

By using this formulation of Bayes’ theorem, statisticians are able to make inferences about some parameter of interest, θ , given the observed data, x . This

density function, $f(\theta|x)$, is referred to as the *posterior distribution* of θ , and it represents a probability density function for the latent variable, θ , the primary quantity of interest, based on the observed data, x . In order to estimate the posterior distribution and make inferences about it, three pieces of information are required: (a) estimation of the function, $f(x|\theta)$, often termed the *likelihood function*, which represents a statistical model that has been fit to the distribution of observed data, x , given the underlying parameter, θ ; (b) estimation of $f(\theta)$, referred to as the *prior distribution*, which represents either the empirical or the expected distribution of the parameter, θ , in the population; and (c) estimation of $f(x)$, which represents the empirical distribution of the observed data, x .

In most practical applications, the use of Bayes' theorem for estimating the posterior distribution is simplified because, for any given data analysis, the empirical distribution of observed data, $f(x)$, will be a constant for any given value of the parameter, θ , and therefore does not need to be included. That is, the observed distribution of data will not affect estimation of θ because $f(x)$ is fixed and does not change. As a result, it is typically dropped from the formula for estimating the posterior distribution:

$$f(\theta|x) \propto f(x|\theta)f(\theta).$$

The preceding formula is read, "the posterior distribution of θ given x , $f(\theta|x)$, is proportional to the likelihood function, $f(x|\theta)$, multiplied by the prior distribution of θ , $f(\theta)$." Bayesian statisticians typically rely on estimating a function that is proportional to $f(\theta|x)$ because estimation of $f(x)$ is unnecessary to estimating the parameters from the posterior distribution. Leaving $f(x)$ out of the formula for the posterior distribution will not affect any of the resulting parameter estimates.

For any analytical application that involves fitting a parametric statistical model to observed data (e.g., an item response model fit to test data, a linear regression model fit to a matrix of continuous variables, or any other analysis that involves the estimation of parameters thought to model the behavior of observed data), analysis proceeds by estimating the likelihood

function, $f(x|\theta)$, and multiplying it by the prior distribution, $f(\theta)$. This produces an estimate that is proportional to the posterior distribution of θ given x , and this distribution is used to make inferences about the parameter, θ .

Inferences from the posterior distribution are typically made by determining point estimates for θ , either by finding the mean of the posterior distribution (referred to as an *expected a posteriori* estimate) or by determining the mode of the posterior distribution (referred to as a *modal a posteriori* estimate). The standard error of θ is determined by estimating the standard deviation of the posterior distribution.

Bayesian statistical methods are distinguished from the more traditional approach (referred to as *Frequentist* methods) in that inferences are made based on the posterior distribution, which cannot be directly observed. Frequentist statisticians typically rely solely on the likelihood function, $f(x|\theta)$, as a basis for making inferences because it represents the model that was actually fit to the observed data. Both methodologies estimate the likelihood function based on observed data, but Bayesian procedures also incorporate information from the prior distribution in order to make inferences. Because of this difference, there exists a philosophical schism between Bayesian and Frequentist statisticians, and it occurs because, generally speaking, the prior distribution for θ is an unobserved probability density function that must be somehow estimated by the analyst. This can often be done by making some reasonable assumptions about the distribution of θ in the population or by collecting data and empirically estimating this function. However, the fact remains that the statistician can never be sure that the particular choice of a prior distribution is accurate and therefore cannot be sure how well the posterior distribution represents the distribution of θ given x . The impact of the prior distribution on the final results of the analysis (i.e., the posterior distribution) will vary depending on the statistician's choice for the distribution. Prior distributions that have significant influence on the posterior distribution are referred to as relatively *informative*, and prior distributions with relatively little influence are called *noninformative* priors.

One cannot know how well suited a particular prior distribution is for estimating a posterior distribution, and so the choice of whether to conduct a Frequentist or a Bayesian data analysis often comes down to philosophical concerns. The perspective of the Bayesian statistician is that by making some reasonable assumptions about the underlying distribution of θ in the population, one can make inferences about the quantities that are of direct interest to the analyst. The perspective of the Frequentist statistician is that one can never appropriately make inferences that go beyond what the data alone suggest. Many statisticians subscribe to the Bayesian school of thought, not only for its intuitive appeal, but also because situations exist in which Bayesian methodologies may be employed where Frequentist analyses are either intractable or impossible. These situations may occur when (a) the likelihood function from a Frequentist analysis is irregular, possibly indicating relatively poor model-data fit, which results in difficulties when determining point estimates and standard errors for θ , or (b) the statistical model is being fit to a relatively sparse data matrix (i.e., there are very few observed data points), making estimation of parameters difficult with Frequentist methods. In both of these situations, Bayesian methods may be employed to produce reasonable parameter estimates.

—William P. Skorupski

See also Posterior Distribution; Prior Distribution

Further Reading

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

BAYLEY SCALES OF INFANT DEVELOPMENT

The Bayley Scales of Infant Development (BSID-II, published by Psychological Corporation) are a set of scales that take approximately 45 minutes to administer

to infants and young children (ages 1 month to 42 months) in order to assess mental, physical, emotional, and social development. The BSID has four main uses: to identify developmental delays (diagnostic tool), to monitor progress after intervention (intervention tool), to teach parents about their child's development (teaching tool), and to compare individual and group differences (research tool).

The BSID is composed of three scales: the Mental Scale, the Motor Scale, and the Behavior Rating Scale. The Mental Scale assesses sensory and perceptual ability, acquisition of object constancy, memory, problem solving, learning, the beginning of verbal communication, and mental mapping. Sample items include discriminating between a bell and a rattle and tracking a moving person with the eyes. The Motor Scale assesses degree of body control, large muscle coordination (sitting and walking), finer manipulation skills of the hands and fingers (picking up small objects), dynamic movement and praxis, and postural imitation. Both the mental and motor scales produce a standardized score. The Behavior Rating Scale assesses attention and arousal, orientation and engagement, and emotional regulation. The Behavior Rating Scale is a 5-point scale (formerly called the Infant Behavior Record) and assesses the developmental level for the status of emotional and social development. The Behavior Rating Scale scores are based on caregivers' reports as well as an examiner's judgments and is completed after the administration of the Mental and Motor scales. This process produces a percentile score for comparison to a nonclinical population.

The test was updated and released in October 2005 by Harcourt Assessment (www.harcourt.com) as the Bayley Scales of Infant and Toddler Development, third edition (Bayley-III). The Bayley-III additions include a Social-Emotional subtest, Adaptive Behavior subtest, Screening Test, Caregiver Report, Scoring Assistant, Growth Scores, and Growth Chart. The Bayley-III uses a current normative sample representing 1,700 children stratified according to age, based on the 2000 U.S. Census. The Bayley-III also has nonnormative data available for children with specific clinical diagnoses, such as autism and Down

syndrome. The BSID is widely used in research settings and has excellent psychometric characteristics.

—Heather Doescher

See also Fagan Test of Infant Intelligence; Intelligence Tests

Further Reading

Black, M. (1999). *Essentials of Bayley Scales of Infant Development II assessment*. New York: Wiley.

Schaefer, E. S., & Bayley, N. (1963). Maternal behavior, child behavior, and their intercorrelations from infancy through adolescence. *Monographs of the Society for Research in Child Development*, 28(3).

Nancy Bayley biography: <http://www.webster.edu/~woolfm/bayley.html>

BECK DEPRESSION INVENTORY

The Beck Depression Inventory (BDI) and the second edition, the Beck Depression Inventory-II (BDI-II), are depression screening instruments published by the Psychological Corporation (www.harcourtassessment.com). The BDI-II is a 21-item self-report instrument (approximate administration time: 5–10 minutes) used to detect and estimate the overall severity of depression in adolescents and adults aged 13 years and older. The instrument can be administered orally as well as in group settings to clinical and normal patient populations. Symptoms of depression are evaluated according to criteria set forth in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.).

The original BDI, developed in 1961, was based on the clinical observations of Dr. Aaron T. Beck and his associates and the typical verbal descriptions reported by depressed psychiatric patients. Representative depressive attitudes and symptoms were consolidated into 21 items read aloud to patients by trained interviewers. The inventory underwent revisions in 1971 at the Center for Cognitive Therapy, University of Pennsylvania, and the amended version, the BDI-IA, was copyrighted in 1978 and published in 1979. In the

original version, respondents were instructed to rate various domains relating to their mood “right now,” whereas in the revised version, instructions asked for mood ratings for the “past week, including today.” The most significant revision of the BDI took place in 1996. This edition, the BDI-II, instructs respondents to provide ratings for the past two weeks. Recently, the BDI-FastScreen, a 7-item self-report measure, has been introduced for use with medical patients.

Each BDI-II item represents a particular symptom of depression: sadness, pessimism, past failure, loss of pleasure, guilty feelings, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, crying, agitation, loss of interest, indecisiveness, worthlessness, loss of energy, changes in sleeping pattern, irritability, changes in appetite, concentration difficulty, tiredness or fatigue, and loss of interest in sex. Four statements in order of severity are presented to the patient for each item and rated on a 4-point scale ranging from 0 to 3. A total depression score is obtained by summing the ratings for the responses to all 21 items. The suggested cutoff scores are 0–13, minimal depression; 14–19, mild; 20–28, moderate; and 29–63, severe.

During the past four decades, the BDI has been used extensively for clinical as well as research purposes and translated into more than 25 languages. In clinical settings, the BDI is often an important component of a comprehensive psychiatric evaluation, and it is used to monitor treatment progress. In empirical studies, the instrument is commonly selected as an outcome measure to demonstrate treatment efficacy. The psychometric characteristics of the BDI-II have been established in groups of college students and psychiatric outpatients.

—Marjan Ghahramanlou-Holloway
and Kathryn Lou

See also Carroll Depression Scale; Clinical Assessment of Depression

Further Reading

Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding

use of the Beck Depression Inventory. *Cognitive Therapy & Research*, 11, 289–299.

Aaron T. Beck Web page: <http://mail.med.upenn.edu/~abeck/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Boothby, J. L., & Durham, T. W. (1999). Screening for depression in prisoners using the Beck Depression Inventory. *Criminal Justice and Behavior*, 26(1), 107–124.

Especially when it comes to working in the field of mental health, using screening tools that are accurate is extremely important. In this study, 1,494 prisoners completed the **Beck Depression Inventory** as part of the admission process to the North Carolina state prison system. The mean score for this population corresponds to the “mild depression” range on the instrument. While overall scores for prisoners were elevated relative to general population norms for the test, female inmates, younger prisoners, close custody inmates, and those serving their first period of incarceration produced even higher Beck scores. Results suggest that a score of 20 might serve as an appropriate cutting score to determine the need for further assessment and mental health intervention in this group. Other analysis of the inmates’ responses yielded four distinct, interpretable factors or groups of variables, labeled (a) cognitive symptoms, (b) vegetative symptoms, (c) emotional symptoms, and (d) feelings of punishment.

BEHAVIOR ASSESSMENT SYSTEM FOR CHILDREN

The *Behavior Assessment System for Children, Second Edition* (BASC-2), published by AGS Publishing (www.agsnet.com), is a set of rating scales and forms that gather information about a child’s behavior, including ratings from parents and teachers as well as children’s self-reports. It is designed to assess and identify children and young adults 2 to 25 years of age with emotional and behavioral disorders. There are five components, which may be used separately or in

combination with one another: (a) a parent rating scale, (b) a teacher rating scale, (c) a self-report scale for the child to describe his or her own emotions and self-perceptions, (d) a structured developmental history form, and (e) a form for recording and classifying classroom behavior. By looking at both positive and negative features, the BASC not only evaluates personality, behavioral problems, and emotional disturbances; it also identifies positive attributes that may help in intervention. Analyzing the child’s behavior from three perspectives—self, teacher, and parent—fosters a comprehensive picture that helps with educational classifications and clinical diagnoses.

The teacher and parent scales gather age-appropriate information on descriptions of observable behavior. The forms describe specific behaviors that are rated on a 4-point scale of frequency, ranging from *never* to *almost always*. Respondents are asked to read the statements and mark the response that best describes how the child has acted over the past 6 months. Teacher and parent forms include statements such as “Adjusts well to changes in routine” and “Annoys others on purpose.” The child’s self-report form consists of 139 statements. For the first 51 items, children choose whether each statement is true or false for them. For the rest of the items, children rate behaviors on the same 4-point scale that the parents and teachers use. The child’s self-report scale includes items such as “I never seem to get anything right” and “I get into trouble for not paying attention.”

The BASC-2 assesses both positive (adaptive) and negative (clinical) dimensions of behavior. When the questionnaire is scored, it provides information about 14 specific areas of a child’s life, which are called scales. Five composite scales provide information about broader aspects of the child’s life by combining the scores from 2 or more of the original 14 scales. Composite scales on the child report include School Problems, Internalizing Problems, Inattention/Hyperactivity, an Emotional Symptoms Index, and Personal Adjustment. High scores indicate higher risk on 10 of the clinical scales and 4 of the composite scales. Low scores indicate higher risk on the remaining 4 adaptive scales and 1 composite scale (Personal Adjustment).

Norms are based on a sample of 13,000 students, ages 2 to 18, from throughout the United States. The BASC-2 is used in both schools and clinics. The test was updated in 2004 by the addition of new scales and the extension of the age range to include college students. The new scales include Functional Communication, Activities of Daily Living, Attention Problems, and Hyperactivity.

—Kristen M. Kalymon

See also Vineland Adaptive Behavior Scales

Further Reading

Reynolds, C. R., & Randy, K. W. (2002). *The clinician's guide to the Behavior Assessment System for Children (BASC)*. New York: Guilford.

BASC-2 information: <http://www.agsnet.com/Group.asp?nGroupInfoID=a30000>

BEHRENS-FISHER TEST

A common question in statistics involves testing for the equality of two population means, μ_1 and μ_2 , based on independent samples. In many applications, it is reasonable to assume that the population variances, σ_1^2 and σ_2^2 , are equal. In this case, the question will usually be addressed by a two-sample t test. The problem of testing for equality of means when the population variances are not assumed to be the same is harder, and is known as the *Behrens-Fisher problem*.

Suppose we have two samples, $x_{11}, x_{12}, \dots, x_{1,n_1}$ and $x_{21}, x_{22}, \dots, x_{2,n_2}$, where the x_{1i} are normally distributed with mean μ_1 and variance σ_1^2 and the x_{2i} are normally distributed with mean μ_2 and variance σ_2^2 , all observations are independent, and it is not assumed that $\sigma_1^2 = \sigma_2^2$. Let $\bar{x}_i$ and s_i^2 denote respectively the mean and variance of sample $i = 1, 2$. Now, $\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ exactly if the original samples are from a normal distribution, and asymptotically if they are not. So, the assumption of normality is not in fact needed.

If we define a pooled variance by

$$s^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2,$$

then, with or without the assumption of normality, s^2 converges to the same quantity, namely a weighted average of σ_1^2 and σ_2^2 ,

$$\sigma_w^2 = \frac{w}{1+w} \sigma_1^2 + \frac{1}{1+w} \sigma_2^2,$$

where w is the limit of the ratio n_1/n_2 and $n_1/n_2 \rightarrow w$ as $n_1, n_2 \rightarrow \infty$.

It can be shown that the usual t statistic,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s \sqrt{1/n_1 + 1/n_2}},$$

instead of converging to $N(0,1)$ under the null hypothesis of no difference in the population means, converges to a normal distribution with mean 0 and variance $(\delta + w)/(\delta w + 1)$, where δ is the ratio between σ_1^2 and σ_2^2 .

In order to understand the effect of the assumption that δ is not necessarily equal to 1 on inference in this setting, it helps to examine how the expression for the asymptotic variance changes as w and δ vary. It is important to realize that if $w = 1$, that is, the two sample sizes are equal, either exactly or in the limit, then the asymptotic variance is 1, no matter the value of δ . Thus, with equal sample sizes, inference, at least asymptotically, is not affected by unequal variances. Having nearly equal samples from the two populations thus mitigates the Behrens-Fisher testing problem. Similarly, if the discrepancies in the population variances are not large, such that $\delta = 1$ or nearly so, then we are back in the standard situation, and again asymptotic inference will proceed as before.

The most worrisome situation is when w is small and δ is large. This corresponds to having a much smaller sample from the first population than from the second, when the variance in the first population is much larger than the variance in the second. In this situation, it is necessary to confront the Behrens-Fisher problem directly. A convenient solution, which is only approximate, is to use Welch's t' statistic, defined as

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which acknowledges the fact that the two sample variances cannot be assumed equal. The difference in the procedure derives from the degrees of freedom associated with this test, since it can be shown that, approximately,

$$\frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \sim \frac{\chi_v^2}{v},$$

where χ_v^2 is a chi-square distribution with v degrees of freedom. The degrees of freedom, v , can be estimated, using *Satterthwaite's approximation*, as

$$\hat{v} = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{s_2^2}{n_2})^2}.$$

To demonstrate that Welch's t' statistic, together with the Satterthwaite approximation for the degrees of freedom, can have an impact on inference, consider the following example, taken from Casella and Berger. Data were collected on wood from a Byzantine church. The samples were from either the core of the church ($n_1 = 14$ measurements) or the periphery ($n_2 = 9$ measurements), and for each, the date of the wood was determined. Is there a difference in the mean age of wood in the core and in the periphery? The data are as follows: core—1294, 1251, 1279, 1248, 1274, 1240, 1264, 1232, 1263, 1220, 1254, 1218, 1251, 1210; periphery—1284, 1274, 1272, 1264, 1256, 1256, 1254, 1250, 1242.

Summary statistics on the two samples are as follows: $\bar{x}_1 = 1249.857$, $s_1^2 = 591.363$, $\bar{x}_2 = 1261.333$, $s_2^2 = 176$. Applying the usual two-sample t test gives $t = -1.29$ on 21 degrees of freedom, which has a p value of 0.21. There is no reason to reject the null hypothesis, and we conclude that there is no significant difference in the age of the wood in the two locations.

Applying Welch's test with the Satterthwaite approximation yields $t' = -1.43$ on 22.3 degrees of

freedom, which has a p value of 0.08. Now the result is borderline significant, by traditional standards, and we might conclude that there is evidence for some difference in age.

Other solutions besides Welch's test have been suggested, including the use of nonparametric statistical tests, resampling (bootstrap), and a Bayesian approach using uniform priors on $(\mu_1, \mu_2, \log \sigma_1, \log \sigma_2)$.

—Nicole Lazar

See also Significance Level; Statistical Significance

Further Reading

- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.
- Miller, R. G. (1986). *Beyond ANOVA: Basics of applied statistics*. New York: Wiley.
- Scheffé, H. (1970). Practical solutions to the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65, 1501–1508.

BENDER VISUAL MOTOR GESTALT TEST

Lauretta Bender's Visual Motor Gestalt Test was developed in 1938 and is commonly referred to as the Bender-Gestalt Test (published by Riverside Publishing Company, www.riversidepublishing.com). It consists of a series of designs printed on individual cards, to be copied by the examinee with pencil and paper. Bender's scoring system evaluated the overall quality of each design and provided an assessment of visual-motor functioning. For comparative purposes, Bender provided graphs and a summary chart of the types of drawings made by children from 3 to 11 years of age. Over the next 65 years, a number of variations in administering and scoring the test emerged to assess visual-motor functioning, psychopathology, and organic brain dysfunction in children and adults. Some of the more prominent variations included scoring systems that examined specific errors (e.g., failure to integrate parts of

designs, rotation of designs, simplification of parts of designs), the use of a background interference procedure during administration of the test (paper containing random lines provided), and a reduction of the number of designs for administration to preschool and early-primary-school children. The test enjoyed considerable success among practitioners and became one of the most widely used tests in psychology.

The revision of the test in 2003 included the addition of several new designs; a memory test; separate tests to assess motor skill and visual perception; a new, easy-to-use scoring system; and a large, nationally representative sample covering the ages 4 to 85+ years. Administration involves the copy phase, followed immediately by the memory phase. In the copy phase, the examinee is instructed to copy each design as it is presented. The memory phase requires the examinee to redraw as many of the designs as possible from memory. Scoring is based on a 5-point scale that assesses the overall quality of each drawing. Standard scores and percentile scores are available for both the copy and the memory phases. If an examinee's scores are low, the supplemental motor and perception tests can be administered to help determine whether the examinee's difficulty is motoric, perceptual, or the integrated visual-motor process.

Research on nonclinical samples as well as a variety of clinical samples, such as individuals with learning disabilities, mental retardation, attention deficit/hyperactivity disorder, autism, and Alzheimer's disease, indicates that the test is a reliable, valid measure of visual-motor functioning and a useful addition to educational, psychological, and neuropsychological test batteries.

—Gary G. Brannigan

Further Reading

- Tolor, A., & Brannigan, G. G. (1980). *Research and clinical applications of the Bender-Gestalt Test*. Springfield, IL: Charles C Thomas.
- Tolor, A., & Schulberg, H. C. (1963). *An evaluation of the Bender-Gestalt Test*. Springfield, IL: Charles C Thomas.

BERNOULLI, JAKOB (1654–1705)

Jakob (James, Jacques) Bernoulli, from Switzerland, was the first of eight members of the mathematically talented Bernoulli family. As directed by his parents, he was trained as a philosopher (master's degree, 1671) and theologian (licentiate, 1676) at the University of Basel.

His career interests, however, were in mathematics and, at least initially, in its application to astronomy. He studied mathematics during his extensive travels, subsequent to graduation, with such luminaries as Nicolas Malebranche (1638–1715) for two years in France, Johann van Waveren Hudde (1628–1704) in the Netherlands, and briefly with Robert Boyle (1627–1691) and Robert Hooke (1635–1703) in England. He then resettled in Basel, and while awaiting a more lucrative offer (and publishing a flawed theory pertaining to comets), he opened a private school of mathematics in 1682. The following year, he was appointed to a teaching post in mechanics at his alma mater, and in 1687, he obtained a professorship and chair in mathematics, which he held until his death.

In 1682, he became a correspondent disciple of Gottfried Wilhelm Leibniz (1646–1716), who co-invented the calculus along with Sir Isaac Newton (1642–1727). This primarily distance-learning arrangement led to Bernoulli's instrumental role in the development of elementary differential and integral calculus and ordinary differential equations.

Among his many discoveries were the system of polar coordinates, the isochrone (the path that an object falls with uniform velocity), and the logarithmic spiral ($r = ae^{b\theta}$). He extended trigonometric functions to complex variables, which led analysis (the study of infinite series) into the study of algebra. Although the $\int$ symbol was invented by his younger brother and student, Johann (1667–1748), the term *integral* was coined by Jakob in an article published in *Acta Eruditorum* in 1690.

His magnum opus, *Ars Conjectandi*, was published posthumously in 1713. Bernoulli spent 20 years writing

the book but never brought it to fruition. It was one of the earliest rigorous treatises on probability theory. In the second of four parts, he proves by induction the binomial theorem. The fourth part contains the *theorem of Bernoulli*. Siméon-Denis Poisson (1781–1840), a descendent in Bernoulli's academic genealogy, renamed the theorem the *law of large numbers*. The modern Monte Carlo method, a technique of repeated sampling, is also known as *Bernoulli trials*.

—Shlomo S. Sawilowsky

See also Pascal, Blaise; Probability Sampling

Further Reading

- Burton, D. M. (1997). *The history of mathematics*. (3rd ed.). New York: McGraw-Hill.
- Sawilowsky, S. S. (2004). A conversation with R. Clifford Blair on the occasion of his retirement. *Journal of Modern Applied Statistical Methods*, 3(2), 518–566.
- Strunk, D. J. (1987). *A concise history of mathematics* (4th rev. ed.). New York: Dover.

BINOMIAL DISTRIBUTION/ BINOMIAL AND SIGN TESTS

The binomial distribution models repeated choices between two alternatives. For example, it will give the probability of obtaining 5 tails when tossing 10 coins or the probability of a rat's choosing 10 times out of 20 the correct branch of a three-branch maze. The binomial test uses the binomial distribution to decide whether the outcome of an experiment using a binary variable (also called a *dichotomy*) can be attributed to a systematic effect. The sign test is applied to before-after designs and uses the binomial test to evaluate whether the direction of change between before and after the treatment is systematic.

Binomial Distribution

The binomial distribution models experiments in which a repeated binary outcome is counted. Each binary outcome is called a *Bernoulli trial*, or simply a

trial. For example, if we toss five coins, each binary outcome corresponds to *H* or *T*, and the outcome of the experiment could count the number of *T* out of these five trials.

Notations and Definitions

We call *Y* the random variable counting the number of outcomes of interest, *N* the total number of trials, *P* the probability of obtaining the outcome of interest on each trial, and *C* a given number of outcomes. For example, if we toss four coins and count the number of heads, *Y* counts the number of heads, *N* = 4, and *P* = ½. If we want to find the probability of getting two heads out of four, then *C* = 2.

With these notations, the probability of obtaining *C* outcomes out of *N* trials is given by the formula

$$\Pr(Y = C) = \binom{N}{C} \times P^C \times (1 - P)^{N-C}. \quad (1)$$

The term $\binom{N}{C}$ gives the number of combinations of *C* elements from an ensemble of *N*; it is called the *binomial of N and C* and is computed as

$$\binom{N}{C} = \frac{N!}{C!(N - C)!} \text{ where } N! = 1 \times 2 \dots \times N. \quad (2)$$

For example, if the probability of obtaining two heads when tossing four coins is computed as

$$\begin{aligned} \Pr(Y = 2) &= \binom{4}{2} \times P^2 \times (1 - P)^{4-2} \\ &= \binom{4}{2} P^2 (1 - P)^{4-2} \\ &= 6 \times .5^2 \times (1 - .5)^2 \\ &= 6 \times .5^4 = .3750, \end{aligned} \quad (3)$$

the mean and standard deviation of the binomial distribution are equal to

$$\mu_Y = N \times P \text{ and } \sigma_Y = \sqrt{N \times P \times (1 - P)}. \quad (4)$$

The binomial distribution converges toward the normal distribution for large values of N (practically, for $P = \frac{1}{2}$ and $N = 20$, the convergence is achieved).

Binomial Test

The binomial test uses the binomial distribution to decide whether the outcome of an experiment in which we count the number of times one of two alternatives has occurred is significant. For example, suppose we ask 10 children to attribute the name “keewee” or “koowoo” to a pair of dolls identical except for their size and that we predict that children will choose keewee for the smaller doll. We found that 9 children out of 10 chose keewee. Can we conclude that children choose systematically? To answer this question, we need to evaluate the probability of obtaining 9 keewees or more than 9 keewees if the children were choosing randomly. If we denoted this probability by p , we find (from Equation 1) that

$$\begin{aligned} p &= \text{Pr}(9 \text{ out of } 10) + \text{Pr}(10 \text{ out of } 10) \\ &= \binom{10}{9} \times P^9 \times (1 - P)^{10-9} + \binom{10}{10} \\ &\quad \times P^{10} \times (1 - P)^0 \\ &= (10 \times .5^9 \times .5^1) + (1 \times .5^{10} \times .5^0) \\ &= .009766 + .000977 \\ &\approx .01074. \end{aligned} \quad (5)$$

Assuming an alpha level of $\alpha = .05$, we can conclude that the children did not answer randomly.

$$P \neq \frac{1}{2}$$

The binomial test can be used with values of P different from $\frac{1}{2}$. For example, the probability p of having five out of six rats choosing the correct door out of four possible doors in a maze uses a value of $P = \frac{1}{4}$ and is equal to

$$\begin{aligned} p &= \text{Pr}(6 \text{ out of } 6) + \text{Pr}(5 \text{ out of } 6) \\ &= \binom{6}{6} \times P^6 \times (1 - P)^{6-6} + \binom{6}{5} \\ &\quad \times P^5 \times (1 - P)^{6-5} \\ &= \frac{1}{4^6} + 6 \times \frac{1}{4^5} \times \frac{3}{4} = \frac{1}{4^6} + \frac{18}{4^6} \\ &\approx .0046. \end{aligned} \quad (6)$$

And we will conclude that the rats are showing a significant preference for the correct door.

Large : Normal Approximation

For large values of N , a normal approximation can be used for the binomial distribution. In this case, p is obtained by first computing a z score. For example, suppose that we had asked the doll question to 86 children and that 76 of them chose keewee. Using Equation 4, we can compute the associated z score as

$$Z_Y = \frac{Y - \mu_Y}{\sigma_Y} = \frac{76 - 43}{4.64} \approx 7.12. \quad (7)$$

Because the probability associated with such a value of Z is smaller than $\alpha = .001$, we can conclude that the children did not answer randomly.

Sign Test

The sign test is used in repeated measures designs that measure a dependent variable on the same observations before and after some treatment. It tests whether the direction of change is random or not. The change is expressed as a binary variable taking the value + if the dependent variable is larger for a given observation after the treatment or – if it is smaller. When there is no change, the change is coded 0 and is ignored in the analysis. For example, suppose that we measure the number of candies eaten on two different days by 15 children and that between these two days, we expose the children to a film showing the danger of eating too much sugar. On the second day, of these

15 children, 5 eat the same number of candies, 9 eat less, and 1 eats more. Can we conclude that the film diminished candy consumption? This problem is equivalent to comparing 9 positive outcomes against 1 negative with $P = 1/2$. From Equation 5, we find that such a result has a p value smaller than $\alpha = .05$, and we conclude that the film did change the behavior of the children.

—Hervé Abdi

Further Reading

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

BINOMIAL TEST

A binomial experiment is one with two outcomes. If one of the outcomes is identified as a success with probability on each trial of π , then the probability of r successes in N trials of the experiment is given by $P(r)$ where

$$P(r) = \frac{N!}{r!(N-r)!} \pi^r (1-\pi)^{N-r}.$$

To test the null hypothesis, $H_0: \pi = \pi_0$, for some constant, $0 < \pi_0 < 1$, against a one-sided alternative requires the summation of all $P(r)$ for all values from r to the desired end point (0 or N). For example, testing the alternative, $H_1: \pi > \pi_0$, we add $P(X)$ for all $X = r, \dots, N$ and define the sum as $p = \sum P(X)$. The result, p , is known as the p value, or exact probability. If $p \leq \alpha$, then the null hypothesis can be rejected at level α . For directional tests in the opposite direction, we take $p = \sum P(X)$, $X = 0, \dots, r$.

For nondirectional tests, the value of p can be calculated in a variety of ways. The definition used here is $p = \sum P(X)$, for all X such that $P(X) \leq P(r)$. If $p \leq \alpha$, the test is significant. The observed success rate, $\pi = r/N$, is significantly different from π_0 at level α .

Suppose a six-sided die is rolled seven times, and the side numbered 5 is defined as a success. Suppose

further that in the seven rolls, there is one roll resulting in a 5. That is, $N = 7$ and $r = 1$. The null hypothesis for a fair die would be $H_0: \pi = 1/6$. For the one-sided alternative, we take $H_1: \pi < 1/6$. Alternatively, we could say the expected number of successes is $\mu = N\pi = 7(1/6) = 1.1667$. In that case, we could express the null hypothesis as $H_0: \mu = 1.1667$.

To test the null hypothesis against the alternative that $\pi < 1/6$, we calculate

$$P(0) = \frac{7!}{0!(7-0)!} \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{7-0}$$

$$P(0) = \left(\frac{5}{6}\right)^7$$

$$P(0) = 0.279082.$$

We also calculate

$$P(1) = \frac{7!}{1!(7-1)!} \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{7-1}$$

$$P(1) = 7 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^6$$

$$P(1) = 0.390714.$$

The one-sided p value becomes $p = 0.279082 + 0.390714 = 0.669796$. Testing at $\alpha = .05$, we do not reject the null hypothesis because $.67 > .05$. That is, the observed rate of side 5 in the rolls is $\hat{\pi} = 1/7 = .143$ and is not significantly less than $1/6 = .167$.

Now suppose on the seven rolls, we obtain side 5 on four of the seven rolls. That is, $N = 7$ and $r = 4$. If we want to test the one-sided hypothesis that the observed rate of $\hat{\pi} = 4/7 = 0.571$ is significantly greater than $1/6 = 0.167$, we need

$$P(4) = 0.015629$$

$$P(5) = 0.001875$$

$$P(6) = 0.000125$$

$$P(7) = 0.000004.$$

In this case we have

$$p = 0.015629 + 0.001875 + 0.000125 + 0.000004 \\ p = 0.017633.$$

Again testing at $\alpha = .05$, we reject H_0 because $.0176 < .05$.

In the two-sided test of $H_0: \pi = 1/6$ with $r = 1$ and $N = 7$, we will have $p = 1.0$ because $P(X) < P(1)$ for all other $0 \leq X \leq 7$. Of course, that test is not significant at $\alpha = .05$.

In the two-sided test of $H_0: \pi = 1/6$ with $r = 4$ and $N = 7$, we will have $p = 0.017633$ just as in the one-sided test because $P(X) > P(4)$ for all $0 \leq X \leq 4$. Of course, that test is significant at $\alpha = .05$. In general, we would expect to find that the one-sided and two-sided tests produce different values of p .

There is a normal approximation to the binomial test, and it can be applied with and without a correction for continuity. It has been known for many years that even the continuity corrected version fails to limit the true Type I error rate to the nominal level unless N is extremely large. It is also well known that the binomial test is more powerful than the normal approximation. Therefore, it is somewhat strange that textbooks published as late as 2005 continue to recommend the normal approximations.

With the wide availability of computer analysis and handheld calculators, there is little excuse for the use of the normal approximations. Anyone who anticipates testing binomial hypotheses should probably obtain a suitable calculator and resort to normal approximations only for data sets too large for the calculator.

—Philip H. Ramsey and Patricia Ramsey

See also Binomial Distribution

Further Reading

Ramsey, P. H., & Ramsey, P. P. (1988). Evaluating the normal approximation to the binomial test. *Journal of Educational Statistics*, 13, 173–182.

BIOINFORMATICS

A single remarkable breakthrough of the 21st century is likely to be biotechnology based on bioinformatics

principles and algorithms. Bioinformatics is advanced by different disciplines. Much scientific, industrial, social, political, economic, and religious activity in upcoming years will be influenced by looming advancements in genetic research. Biostatisticians and computational biologists engaged in bioinformatics are working to clearly comprehend how molecular machinery works, fails, and can be repaired. One needs an excellent command of and expertise in biology, calculus, probability, mathematical statistics, and computer science to follow and make contributions in bioinformatics, an emerging discipline that analyzes large genetic data sets using statistical and information techniques. The discipline is growing quickly as a result of the rapid availability of DNA or protein sequence data on the World Wide Web. Because the biological machine is chance oriented, both probability and statistics are fundamental to understanding DNA or protein sequences.

Bioinformatics is one of three branches in a new discipline. The other two branches are medical informatics and health informatics. Medical informatics concentrates on computational algorithms to improve communication and understanding in order to manage medical knowledge and application. Microarray technology is the driving engine of this discipline. Health informatics studies the dynamics among (a) computers, communications, and other information sciences; (b) engineering, technology, and other sciences; and (c) medical research, education, and practice. Bioinformatics is a collection of tools and ideas for deciphering the complexity of molecular machinery. According to bioinformatics, biology is informational science, and this complex and diversified field is increasingly becoming a cross-disciplinary science. It is in its infancy but evolving rapidly. Biostatisticians, computer scientists, operations researchers, and molecular biologists work hard to enrich bioinformatics.

Since the discovery of the helix structure of DNA by James D. Watson and Francis H. C. Crick, several array-based biotechnologies have been constructed to determine and exploit gene expression levels and their interactions. Gene expression is a basic link between genotype and phenotype. Gene expression data are generated on a massive scale. New statistical principles and computing techniques are necessary to meet

the demand for quick and correct interpretations of so much data.

As the territory of bioinformatics is changing dramatically, statisticians have to learn the language and jargon of bioinformatics. For example, much of the so-called simple random sampling, stratifications, randomization, replication, and so on, of the 20th century has become obsolete in the genetic research arena of the 21st century. DNA-oriented research ideas are geared to statistics' being an exact science.

John Naisbitt states in *Megatrends* that "we are drowning in information but starved of knowledge." Fast-improving computing facilities change the way knowledge, discovery, and application in all scientific and day-to-day life are done. Before, genetic data were analyzed using a hypothesis-driven-reductions approach, but now, it is all done by a data-driven approach. Consequently, bioinformatics ideas play a significant role in genetic research.

Bioinformatics is all about identifying genes in genome sequences, figuring out closeness of one sequence to another, and answering questions such as the following: How similar are two different organisms? Where in DNA is a particular gene? What proteins are produced by a particular gene? What are the interrelations between genes and proteins? How are one person's genes different from those of another individual? And how can we design a way to store, process, and analyze this knowledge? Molecular human biology can be summarized as follows: There are 22 chromosomes in paired style. Every human female has two X chromosomes whereas a human male has one X and one Y chromosome. Each chromosome has a single double stranded DNA molecule with complementary nucleotides (A-T, G-C) forming pairs in the strands. The nucleotides are A for adenine, T for thymine, G for guanine, and C for cytosine. There may be redundant information in each strand. Organisms need to produce proteins for a variety of functions in life. There is a code for the start and end of the proteins. Major terms in bioinformatics include *exon* (segment of DNA that supplies information to make proteins), *intron* (a noncoding segment that interrupts exons to produce a proper copy of RNA), and *splice site* (the boundary of an exon and an intron). This site allows the uninterrupted gene or

amino acid sequence of proteins. *Promoter sites* are segments of DNA that start the transcription of genes, enhancing controls of the transcription.

Why should one study bioinformatics? This emerging field seeks to understand the secrets of life's machinery and therefore should be useful in discovering new drugs, custom suited for each patient, to treat illnesses now considered incurable. The complex process of life can perhaps be explained by simple principles of genes!

The Human Genome Project is closely involved with the development of bioinformatics. The project started in 1980 to determine, for medical purposes, patterns in the entire sequence of the 3 million human nucleotides. The draft sequence was completed on October 7, 2000, and was published on February 15, 2001. The sequences of 908 species, including 557 viruses, 112 bacteria, and 172 eukaryotes, have been completed. The human sequence is the largest and was completed on April 25, 2003, the 50th anniversary of the discovery of chromosomes.

The methodologies that are used in bioinformatics may be grouped as follows: The symbolic computations, hidden Markov models, and PERL programming are computer intensive. Another group, consisting of artificial intelligence (using the human paradigm), statistical inference (inductive or deductive), knowledge representation, expert systems, rule-based neural networks, natural languages, pattern discovery, matching machine learning, and hierarchical clustering, is statistical. The probabilistic group consists of decision trees and operations research methods including dynamic programming, probability ideas, information and entropy ideas.

Computer programs such as BAMBE, BLAST, BLAT, FASTA, MEGA, PSI BLAST, VISTA, and VAST do microarray analysis, describe chromosomes, try contig mapping, and explore DNA-RNA databases like EMBL and GENBANK. The hidden Markov models, knowledge discovery, mutations, machine learning methods, neural networks, protein databases, x-y chromosomes, and Zipf's law, among others, are heavy but powerful tools of bioinformatics.

Genetic and molecular epidemiology is evolving into the mainstream of clinical health research. The proliferation of genetic data highlights the importance

of analytical and graphical tools. Through an understanding of the genetic architecture of complex diseases, new modern medicines can be developed. Bioinformaticians are expected to play a critical role in the process of discovering new genetics-based medicines.

Several Monte Carlo techniques are used in bioinformatics. Being interdisciplinary, Monte Carlo techniques attract attention from bioinformaticians, DNA researchers, computer scientists, probabilists, and statisticians. Monte Carlo is a computer-assisted problem-solving technique in a complex system. Its methodologies were used first in physics but are now used widely in bioinformatics. Probabilists and statisticians might trace its history all the way back to the Buffon Needle problem in 1777. It is easy to compute that the chances of a needle of length l intersecting one of the parallel lines separated by a distance of D units is $2l/\pi D$. After throwing the needle a sufficiently large number of n times and letting $p_n = \#$ times intersected / n , the value of π can be accurately approximated, and that is a Monte Carlo technique. That is,

$$\hat{\pi} = \lim_{n \rightarrow \infty} \frac{2l}{p_n D}.$$

In this genomic age, investigations of complex diseases require genetic concepts for understanding and discussing susceptibility, severity, drug efficacy, and drug side effects. In this process, researchers end up analyzing huge amounts of data. Concepts and tools such as multiple testing and data dredging are valuable in the pursuit of such huge data analyses.

The subtopics of bioinformatics are population genetics, evolutionary genetics, genetic epidemiology, animal and plant genetics, probability theory, several discrete and continuous distributions, moment/probability generating functions, Chebychev's inequality, entropy, correlation, distribution of maximum and

minimum in a set of random quantities, Bayesian and classical inference procedures, stochastic processes, Markov chains, hidden Markov models, computationally intensive methods in statistics, shotgun sequencing, DNA models, r -scans, nucleotide probabilities, alignments, dynamic programming, linear gap models, protein sequences, substitution matrices, edge effects in unaligned sequences, both discrete and continuous time evolutionary Jukes-Cantor models, Kimura neutral models, Felsenstein models, and phylogenetic tree estimations in biology, among others.

BLAST (Basic Local Alignment Search Tool), one of the bioinformatics software programs mentioned above, comes in several versions, BLASTP, BLASTN, and BLASTX, for comparing protein sequences, for nucleotide sequences, and for translated sequences, respectively. All BLAST programs produce similar output, consisting of a program introduction, a schematic distribution of alignments of the

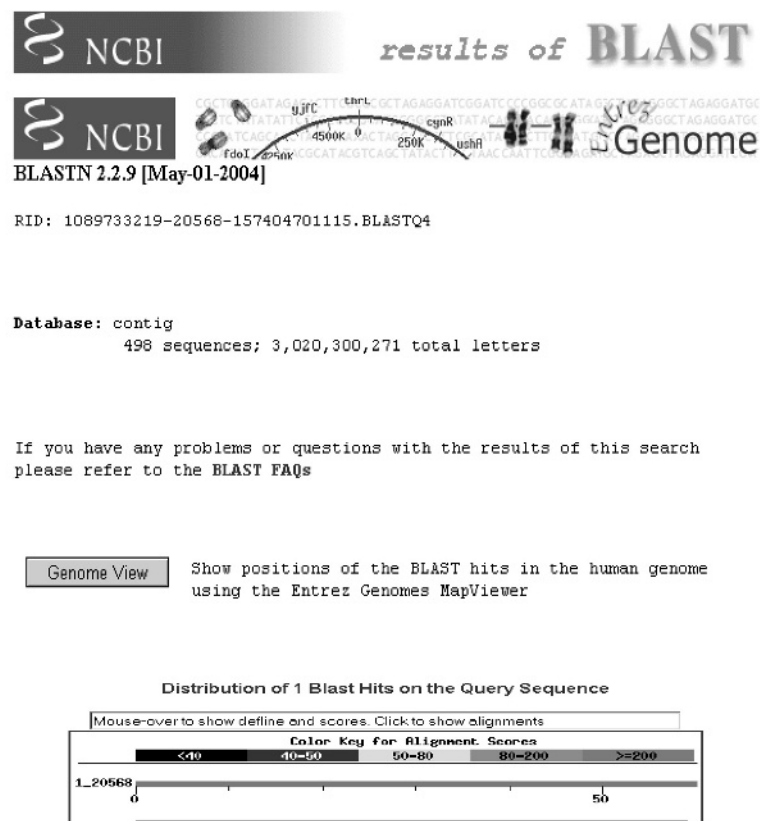


Figure 1 BLAST

Source: www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html.

query sequence to those in the databases, a series of one-line descriptions of the database sequences that have significantly aligned to the query sequence, the actual sequence alignments, and a list of statistics specific to the BLAST search method and version number. (See Figure 1.)

The top 100 significant alignments of the query sequence to database sequences are displayed schematically against the query sequence. Colored bars are distributed to reflect the region of alignment onto the query sequence. The color legend represents alignment scores, the higher scores being most significant. Selecting a bar will cause a description of that specific database sequence to be displayed in the window and allow the browser to jump down to that particular alignment for viewing (Figure 2).

FASTA (abbreviation of *fast algorithms*) performs a fast alignment of all protein/nucleotides sequences. This computer program is based on ideas found in Pearson and Lipman (1988). FASTA searches for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein).

DAMBE (Data Analysis in Molecular Biology and Evolution) can be used for analyzing molecular

sequence data. Although science has unlocked several secrets of life, mysteries remain. For example, how do chromosomes organize themselves in meiosis and mitosis? What are the properties of protein value? How do DNA strands wind up? How do genes transmit instructions to make a specific protein? Why do shorter chromosomal arms have higher rates of recombination? Why is recombination less frequent near the centromeres? Why do more recombinations occur during meiosis? Why do chromosomes 13, 18, and 21 have the fewest genes per kilobase? Why is genome size not correlated with an organism's complexity? Why is only 5% of RNA coding while more than 50% of the repeat sequences are not coding? Why do more A-T than G-C pairings occur in general but not in chromosome 19? Why does the genetic material on the Y chromosome remain relatively young? Why is the mutation rate of the Y chromosome 1.57 times greater than that of the X chromosome? Why do thousands of genes produce noncoding RNA, tRNA, and rRNA? Why do more than 25% of tRNA genes exist on chromosome 6? And what are the functions of many proteins?

There are unresolved cloning and ethical issues. People are divided into those who argue in favor and those who are against cloning and stem cell research. Those in favor of doing research work cite forensic issues, finding cures for certain human diseases, advantages of DNA repair in diabetes and other illnesses, and advantages of understanding heritability. Those who are opposed cite security, privacy, ethics, and the fear of making "Frankenstein's monster."

—Ramalingam Shanmugam

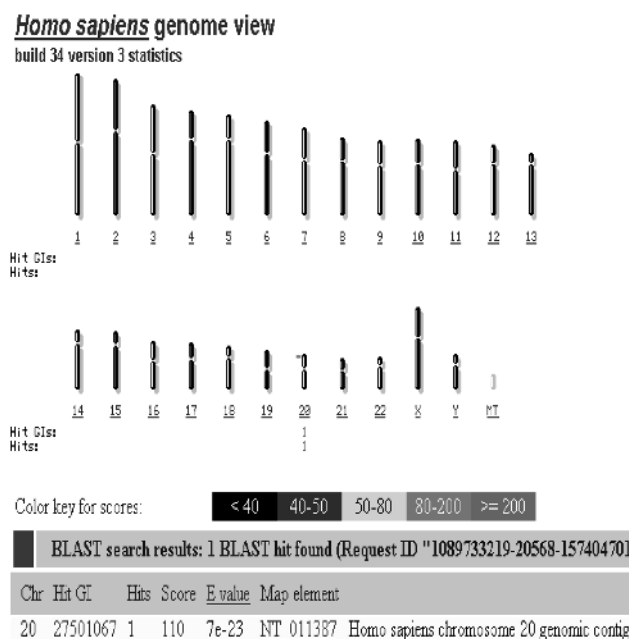


Figure 2 Human Chromosomes

Source: http://www.ensembl.org/Homo_sapiens/mapview.

Further Reading

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of USA National Academy of Sciences*, 85(8), 2444–2448. Retrieved from [http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+\[MEDLINE-pmid:'3162770'\]](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+[MEDLINE-pmid:'3162770'])

Xia, X. (2000). *Data analysis in molecular biology and evolution*. Boston: Kluwer Academic.

Xia, X., & Xie, Z. (2001). DAMBE: Data analysis in molecular biology and evolution. *Journal of Heredity*, 92, 371–373.

BLAST resources: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html>, <http://www.ncbi.nlm.nih.gov/>

DAMBE Windows95/98/NT executables: <http://aix1.uottawa.ca/~xxia/software/software.htm>

National Center for Biotechnology Information Human Genome Resources: <http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/>

National Center for Biotechnology Information Social Analysis of Gene Expression Tag to Gene Mapping: www.ncbi.nlm.nih.gov/SAGE/

National Center for Biotechnology Information Statistics of Sequence Similarity Scores: www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Humphreys, K., Demetriou, G., & Gaizauskas, R. (2000). Bioinformatics applications of information extraction from scientific journal articles. *Journal of Information Science*, 26(2), 75–85.

Information extraction technology developed through the U.S. Defense Advanced Research Projects Agency (DARPA) Message Understanding Conferences (MUCs) has proved successful at extracting information from newswire texts and in domains concerned with human activity. This paper considers the application of this technology to the extraction of information from scientific journal papers in the area of molecular biology. In particular, it describes how an information extraction designed to participate in the MUC exercises has been modified for two **bioinformatics** applications, one concerned with enzyme and metabolic pathways, the other with protein structure. Progress to date provides convincing grounds for believing that information extraction techniques will deliver novel and effective ways for scientists to make use of the core literature that defines their disciplines.

BISERIAL CORRELATION COEFFICIENTS

Biserial correlation coefficients are measures of association that apply when one of the observed variables

takes on two numerical values (a binary variable) and the other variable is a measurement or a score. There are several biserial coefficients, with the appropriate choice depending on the underlying statistical model for the data. The point biserial correlation and Pearson's biserial correlation are arguably the most well known and most commonly used coefficients in practice. We will focus on these two coefficients but will discuss other approaches.

Karl Pearson developed the sample biserial correlation coefficient in the early 1900s to estimate the correlation ρ_{YZ} between two measurements Z and Y when Z is not directly observed. Instead of Z , data are collected on a binary variable X with $X = 0$ if Z falls below a threshold level and $X = 1$ otherwise. The numerical values assigned to X do not matter provided the smaller value identifies when Z is below the threshold. In many settings, the latent variable Z is a conceptual construct and not measurable. The sample point biserial correlation estimates the correlation ρ_{YX} between Y and a binary variable X without reference to an underlying latent variable Z .

We will use S. Karelitz and colleagues' data on 38 infants to illustrate these ideas. A listing of the data is given in Table 1. The response Y is a child's IQ score at age 3, whereas $X = 1$ if the child's speech developmental level at age 3 is high, and $X = 0$ otherwise. The (population) biserial correlation ρ_{YZ} is a reasonable measure of association when X is a surrogate for a latent continuum Z of speech levels. The (population) point biserial correlation ρ_{YX} is more relevant when the relationship between IQ and the underlying Z scale is not of interest, or the latent scale could not be justified.

The Point Biserial Correlation

Assume that a random sample $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ of n observations is selected from the (Y, X) population, where Y is continuous and X is binary. Let s_{YX} be the sample covariance between all y_i and all x_i , and let s_y^2 and s_x^2 be the sample variances of all y_i and all x_i , respectively. The population correlation ρ_{YX} between Y and X is estimated by the sample point biserial correlation coefficient, which is just the

Table 1 Data for a Sample of 38 Children

X = 0	Y:	87	90	94	94	97	103	103	104	106	108	109
		109	109	112	119	132						
X = 1	Y:	100	103	103	106	112	113	114	114	118	119	120
		120	124	133	135	135	136	141	155	157	159	162

Note: X = speech developmental level (0 = low; 1 = high), and Y = IQ score.

product-moment correlation between the Y and X samples:

$$r_{YX} = \frac{s_{YX}}{s_Y s_X}.$$

The sample point biserial estimator r_{YX} can also be expressed as

$$r_{YX} = \frac{(\bar{y}_1 - \bar{y}_0)}{s_Y} \{ \hat{p}(1 - \hat{p}) \}^{1/2},$$

where

$\bar{y}_1$ and $\bar{y}_0$ are the average y values from sampled pairs with $x_i = 1$ and $x_i = 0$, respectively, and

$\hat{p}$ is the proportion of observations that have $x_i = 1$.

The equivalence between the two expressions for r_{YX} requires that the sample variances and covariances be computed using a divisor of n and not the usual divisor of $n - 1$.

The first careful analysis of the properties of r_{YX} was provided by Robert Tate in the middle 1950s. He derived the large-sample distribution of r_{YX} assuming that the conditional distributions of Y given $X = 1$ and given $X = 0$ are normal with potentially different means but the same variances. Tate showed that $T = (n - 2)^{1/2} r_{YX} / (1 - r_{YX}^2)^{1/2}$ is equal to the usual two-sample Student t statistic for comparing $\bar{y}_1$ to $\bar{y}_0$ and that the hypothesis $\rho_{YX} = 0$ can be tested using the p value from the two-sample t test. For $\rho_{YX} \neq 0$, large-sample hypothesis tests and confidence intervals can be based on a normal approximation to r_{YX} , with estimated variance

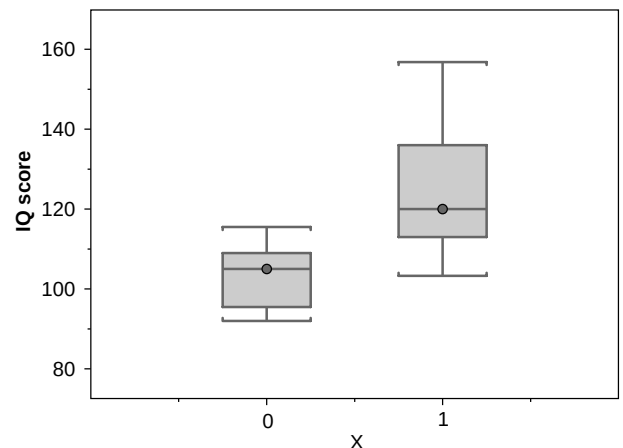
$$\widehat{\text{var}}(r_{YX}) = \frac{(1 - r_{YX}^2)^2}{n} \left\{ 1 - 1.5r_{YX}^2 + \frac{.25r_{YX}^2}{\hat{p}(1 - \hat{p})} \right\}.$$

The biserial estimate r_{YX} is robust in the sense that the bias in r_{YX} approaches 0 as the sample size increases,

even if the distributional assumptions are not satisfied. However, the estimated variance of r_{YX} is sensitive to the assumption of equal variances for the two subpopulations ($X = 0$ and $X = 1$). Somesh Das Gupta generalized Tate's distributional results to allow unequal variances and nonnormal distributions.

Figure 1 gives side-by-side box plots of the IQ data generated by the S-PLUS statistics package. Although the distributions of the IQ scores for the samples with $X = 0$ and $X = 1$ are slightly skewed, the assumptions for Tate's analysis seem plausible. The mean IQ scores in the two samples are $\bar{y}_1 = 2779/22 = 126.32$ and $\bar{y}_0 = 1676/16 = 104.75$. Also, $\hat{p} = 22/38 = 0.579$ and $s_Y = 19.126$, which gives $r_{YX} = 0.557$ and $\widehat{\text{sd}}(r_{YX}) = 0.013$. The large difference between the means of the two groups relative to the within-group spreads is consistent with the observed correlations being significantly different from 0 ($T = 4.024$ on $38 - 2 = 36$ df; p value $< .001$).

A shortcoming of the population point biserial correlation as a measure of association is that ρ_{YX} cannot assume all values between -1 and 1 . The limits on ρ_{YX} depend on the distribution of Y and the $\Pr(X = 1)$. For example, if Y is normally distributed, then $-.798 \leq \rho_{YX} \leq .798$ regardless of $\Pr(X = 1)$. The maximum value can be achieved only when $\Pr(X = 1) = .50$. If Y is normal and $\Pr(X = 1) = .85$, then $-.653 \leq \rho_{YX} \leq .653$. Such restrictions can lead to a misinterpretation of the strength of the sample point biserial correlation. W. Joe Shih and W. H. Huang examined this issue and

**Figure 1** Box plots of IQ Score by X

proposed a way to calibrate the point biserial coefficient.

Pearson's Biserial Correlation

Suppose X is a binary variable that results from categorizing a continuous latent variable Z . Set $X = 1$ when $Z > \theta$ and $X = 0$ otherwise, where θ is a fixed but unknown threshold. Without loss of generality, we can assume that Z is standardized to have mean 0 and variance 1. Let $f(t)$ be the probability density function of Z and note that θ is the upper p th percentile of the distribution of Z ; that is,

$$p = \Pr(X = 1) = \Pr(Z > \theta) = \int_{\theta}^{\infty} f(t)dt.$$

If the conditional expectation of Y given Z is a linear function of Z , then the population biserial correlation and the point biserial correlation are related by

$$\rho_{YZ} = \rho_{YX} \frac{\{p(1-p)\}^{1/2}}{\lambda(\theta, f)},$$

where $\lambda(\theta, f) = \int_{\theta}^{\infty} tf(t)dt$. The linearity assumption is satisfied when (Y, Z) has a bivariate normal distribution, a common assumption, but it holds for other elliptically symmetrical bivariate distributions as well. Under normality,

$$f(t) = \phi(t) \equiv \frac{1}{\sqrt{2\pi}} \exp(-.5t^2),$$

$$\lambda(\theta, f) = \phi(\theta), \text{ and}$$

$$\rho_{YZ} = \rho_{YX} \frac{\{p(1-p)\}^{1/2}}{\phi(\theta)}.$$

The relationship between the population biserial correlation ρ_{YZ} and ρ_{YX} can be exploited to estimate ρ_{YZ} from a random sample $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ when the distribution of Z is known. Edward J. Bedrick suggested the simple method-of-moments estimator

$$\tilde{r}_{YZ} = r_{YX} \frac{\{\hat{p}(1-\hat{p})\}^{1/2}}{\lambda(\hat{\theta}, f)},$$

where $\hat{\theta}$ is the estimated threshold based on the proportion $\hat{p}$ of sampled pairs with $x_i = 1$ (i.e., $\hat{\theta}$ satisfies $\hat{p} = \Pr(Z > \hat{\theta})$). If Z has a normal distribution, $\tilde{r}_{YZ}$ is Pearson's sample biserial estimator, and

$$\begin{aligned} r_{pb} &= \frac{r_{YX}}{\phi(\hat{\theta})} \{\hat{p}(1-\hat{p})\}^{1/2} \\ &= \frac{(\bar{y}_1 - \bar{y}_0)}{s_Y \phi(\hat{\theta})} \hat{p}(1-\hat{p}). \end{aligned}$$

Bedrick's derivation of $\tilde{r}_{YZ}$ parallels Pearson's original treatment of the biserial correlation coefficient, so $\tilde{r}_{YZ}$ and r_{pb} have similar distributional properties. Bedrick showed that the large-sample distribution of $\tilde{r}_{YZ}$ is normal with mean ρ_{YZ} and gave an expression for the large-sample variance of $\tilde{r}_{YZ}$. In the early 1900s, H. E. Soper gave an estimator for $\text{var}(r_{pb})$ when (Y, Z) is normal:

$$\begin{aligned} \widehat{\text{var}}(r_{pb}) &= \frac{1}{n} \left[r_{pb}^4 + \frac{r_{pb}^2}{\phi^2(\hat{\theta})} \{ \hat{p}(1-\hat{p})\hat{\theta}^2 \right. \\ &\quad \left. + (2\hat{p}-1)\hat{\theta}\phi(\hat{\theta}) - 2.5\phi^2(\hat{\theta}) \right. \\ &\quad \left. + \frac{\hat{p}(1-\hat{p})}{\phi^2(\hat{\theta})} \right]. \end{aligned}$$

Tate showed that $r_{pb} \geq \sqrt{.5\pi} r_{YX}$ (approximately $1.25r_{YX}$). For the IQ data, $r_{pb} = .703$ and $\widehat{\text{sd}}(r_{pb}) = .133$.

The derivations of r_{pb} and $\tilde{r}_{YZ}$ rely heavily on distributional assumptions. Neither estimate is consistent if either the conditional expectation of Y given Z is not linear in Z or the distribution of Z is incorrectly specified. This lack of robustness is problematic because neither assumption can be checked empirically. Another undesirable property of r_{pb} and $\tilde{r}_{YZ}$ is that the magnitude of these estimates can exceed 1. This anomaly is common in small samples when the population correlation is large but becomes less likely as n increases.

Alternative Estimators

Several researchers developed estimators that eliminate one or more limitations of r_{pb} . Hubert Brogden,

William Clemans, and Frederic Lord generalized Pearson's biserial estimator by relaxing the assumption that the distribution of Z is known. The Clemans-Lord estimator has the attractive property of being bounded between -1 and $+1$ and thus is much less variable than r_{pb} when the population correlation is large. Edward Cureton and Eugene Glass proposed rank-based versions of the biserial estimator.

Tate proposed a maximum likelihood estimator of ρ_{YZ} assuming that (Y, Z) has a bivariate normal distribution. The maximum likelihood estimator is efficient when the model holds, but the estimate is not robust to misspecification of the model and requires specialized software to compute. The maximum likelihood approach could be considered with other probability models. However, Bedrick showed that the large-sample variances of the (noniterative) Clemans-Lord estimator and the maximum likelihood estimator are often close in a variety of normal and nonnormal populations.

—Edward J. Bedrick

Further Reading

- Bedrick, E. J. (1992). A comparison of modified and generalized sample biserial correlation estimators. *Psychometrika*, 57, 183–201.
- Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective inefficiency. *Psychometrika*, 14, 169–182.
- Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, 21, 287–290.
- Das Gupta, S. (1960). Point biserial correlation and its generalization. *Psychometrika*, 25, 393–408.
- Karelitz, S., Fisichelli, V. R., Costa, J., Karelitz, R., & Rosenfeld, L. (1964). Relation of crying activity in early infancy to speech and intellectual development at age three years. *Child Development*, 35, 769–777.
- Lord, F. M. (1963). Biserial estimates of correlation. *Psychometrika*, 28, 81–85.
- Pearson, K. (1909). On a new method of determining the correlation between a measured character A and a character B. *Biometrika*, 7, 96–105.
- Shih, W. J., & Huang, W.-H. (1992). Evaluating correlation with proper bounds. *Biometrics*, 48, 1207–1213.
- Tate, R. F. (1954). Correlation between a discrete and a continuous variable: Point-biserial correlation. *Annals of Mathematical Statistics*, 25, 603–607.

Tate, R. F. (1955). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 42, 205–216.

SAS macro for computing the point, Pearson, and rank biserial coefficients: <http://ftp.sas.com/techsup/download/stat/biserial.html>

BIVARIATE DISTRIBUTIONS

Cause and effect between two phenomena in a real-life setting cannot be judged or resolved without configuring their patterns of occurrence, correlation, and uncertainties. Item response theory has been well developed by psychologists to explain the personal ability level of the examinees who answer a series of questions varying in toughness level. The ability and toughness levels are correlated random variables with a degree of dependence between them. Another revealing example is that the safety of a building cannot be ascertained without knowledge of both stress and strength of the materials used in the building. When the strength, Y , exceeds stress level, X , safety is guaranteed. Another serious life-and-death example is the distance of a populous city from the geological epicenter of an earthquake and the severity of damage in the city, as experienced in the tsunami of December 26, 2004, in south Asia. In a health application, X and Y are “susceptibility” and “immunity” levels of a person in the outbreak of a disease epidemic, respectively, and a person is healthy only so long as Y exceeds X . Identification of underlying bivariate probability distribution in these and other applications reveals a volume of related knowledge.

Such two stochastic aspects x and y in experimental or randomly observed studies are well explained by employing an appropriate underlying joint (probability) distribution $f(x, y)$. Their patterns of occurrence, correlation, and the prediction of one aspect using the occurrence level of another aspect are feasible from randomly collected bivariate data. Though the domain for the data could be a shrunk version depending on the cases, it is in general from minus infinity to positive infinity. A truncated or censored version of the bivariate (probability) distributions

might be employed in such scenarios. A bivariate distribution could be a count of continuous or mixed type. However, their conditional $f(x | Y = y)$ and $f(y | X = x)$ distributions reveal interinfluence by one on another, but their marginal distribution $f(x)$ or $f(y)$ does not. For example, the predicted value of Y for a given level $X = x$ is called a regression function of x . The conditional and marginal dispersions obey an inequality $\text{Var}(Y | X = x) \leq \text{Var}(Y)$, which means that the conditional prediction of Y with knowledge of X is more precise than unconditional prediction of Y . The inverse of variance is called precision. Also, the so-called product moment is built in a hierarchical manner in accordance with the result $E(YX) = E[E(Y | X)]$, where the outer expectation is with respect to the random variable, X . Their covariance is defined to be $\text{cov}(Y, X) = E[E(Y | X = x) - E(Y) E(X)]$. The covariance is scale oriented, and it could be misleading unless caution is exercised. Furthermore, the variance can also be hierarchically constructed according to a result $\text{Var}[Y] = E[\text{Var}(Y | X = x)] + \text{Var}[EY | X = x]$.

As done in univariate cases, the moment, cumulant, and probability generating functions are derived and used to identify central and noncentral moments and cumulants, along with their properties in bivariate distributions. The correlation coefficient ρ_{yx} between designated dependent variable Y and chosen independent (more often called *predictor*) variable X is $\text{cov}(Y, X) / (\sigma_Y \sigma_X)$, where $\sigma_Y = \sqrt{\text{var}(Y)}$ and $\sigma_X = \sqrt{\text{var}(X)}$ are standard deviations. The correlation coefficient is scale free. A simple linear regression function is $Y = \beta_0 + \beta_1 x + \varepsilon$ for predicting Y at a selected level $X = x$, and the so-called regression parameter (slope) is $\beta = \frac{\rho_{yx}}{\sigma_X}$. See a variety of popular bivariate distributions in Tables 1 and 2.

The uncorrelated variables Y and X are clearly independent in the case of bivariate Gaussian random variables but not necessarily in other cases. Two variates are uncorrelated if the joint probability distribution is the product of their marginal probability distributions. Equivalently, the conditional probability distribution is equal to the marginal probability distribution. In a collection of bivariate random samples, if Y_i and X_i in an i^{th} pair are independent, then $Y_{\max}$ and $X_{\max}$ are also independent. The converse is not always true. So Geoffroy proved that if the ratio

$$\frac{1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)}{1 - F_{X,Y}(x, y)}$$

asymptotically joins the horizontal axis, then the independence of $Y_{\max}$ and $X_{\max}$ is sufficient for the independence of the random observations on Y and X where $F(\cdot)$ is a cumulative distribution. However, bivariate Gaussian, logistic, Gumbel, and several other distributions of the type

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) [1 + \alpha (1 - F_X(x)) (1 - F_Y(y))]$$

validate the above ratio condition, independence between $Y_{\max}$ and $X_{\max}$. Popularly employed bivariate distributions for continuous data analysis are displayed in Table 1.

Bayesians view nonobservable *latent* parameter θ as a random variable. Its marginal distribution $p(\theta)$ in its admissible domain $-\infty < \theta < \infty$ is called prior distribution. The conditional distribution

$$p(\theta | y) = \frac{f(y, \theta)}{m(y)} = \frac{p(\theta)l(y|\theta)}{\int_{-\infty}^{\infty} f(y, \theta)d\theta}$$

is called *posterior* (an update of the term *prior*) distribution, where $m(y)$ and $l(y|\theta)$ are called marginal distribution and likelihood function, respectively. The posterior distribution is an update of knowledge on parameter θ based on evidence in data, y . For an example, with known data variance σ_y^2 , the univariate Gaussian distribution $f(y, \theta | \sigma_y^2) = [2\pi\sigma_y^2]^{-1/2} \exp[-(y - \theta)^2 / 2\sigma_y^2]$ is considered as a bivariate distribution of random variables Y and Θ . Note the marginal mean $E(Y) = \theta$ is stochastic and follows independently a prior Gaussian distribution:

$$f(\theta | m, \sigma_0^2) = [2\pi\sigma_0^2]^{-1/2} \exp[-(y-m)^2 / 2\sigma_0^2], \quad -\infty < \theta < \infty, \\ \sigma_0^2 > 0, \quad -\infty < m < \infty.$$

Then, the posterior distribution $f(\theta | y)$ follows a Gaussian distribution with a weighted mean

$$E(\theta | y) = \frac{\frac{1}{\sigma_0^2} y + \frac{1}{\sigma_y^2} m}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2}}$$

and variance $\text{var}(\theta | y)$.

$$\text{var}(\theta | y) = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2} \right)^{-1}.$$

Note that the posterior mean is the weighted average of prior mean m and new data y , where the weights are prior precision and data precision. The precisions moderate prior mean and current data. The importance of the Bayesian style of thinking could not be better advocated by anything other than the concept of bivariate distribution.

Other distributions commonly employed for count data are bivariate versions of binomial, Poisson, inverse binomial, Neyman, Hermite, and logarithmic series distributions.

Consider that a bivariate distribution for the random number of traffic accidents and number of fatalities needs to be selected. The number of accidents Y and number of fatalities X in a specified location during a day can be modeled successfully using a quasi-binomial-Poisson bivariate distribution, as cited in Table 2. The marginal probability distribution of Y is

$$p(Y = y) = \sum_{x=0}^y p(y, x) = \lambda_1 (\lambda_1 + \lambda_2 y)^{y-1} \exp[-(\lambda_1 + \lambda_2 y)] / y!,$$

a quasi-Poisson distribution, where $y = 0, 1, 2, \dots, \infty$, plus the parameters $\lambda_1 > 0$ and $-1 < \lambda_2 < 1$, denote, respectively, the observation space, accident rate, and hazard level. The marginal mean and variance of Y are

$$\frac{\lambda_1}{1 - \lambda_2} \quad \text{and} \quad \frac{\lambda_1}{(1 - \lambda_2)^3},$$

respectively. When there is no accident proneness (that is, $\lambda_2 = 0$), the above marginal probability mass function reduces to the popular Poisson distribution $p(y) = e^{-\lambda_1} \lambda_1^y / y!$. Not all accidents turn out to be fatal. An accident turns out to be a fatal accident with a probability $0 < p < 1$ due to some uncertain causes. The total number of fatal accidents $X = X_1 + X_2 + \dots + X_y$ for a given $Y = y$ follows conditionally a quasi-binomial distribution

$$\begin{aligned} p(X = x | Y = y) \\ &= \binom{y}{x} \left(\frac{\lambda_1 p [1 - p]}{\lambda_1 + \lambda_2 y} \right) \left(\frac{\lambda_1 p + \lambda_2 x}{\lambda_1 p + \lambda_2 y} \right)^{x-1} \\ &\quad \left(\frac{\lambda_1 [1 - p] + \lambda_2 [y - x]}{\lambda_1 p + \lambda_2 y} \right)^{y-x-1}, \end{aligned}$$

where $x = 0, 1, 2, \dots, y$ and $0 < p < 1$ denote, respectively, the observation space for the random number of fatal accidents and the fatality parameter. In this scenario, the number of accidents splits into fatal and nonfatal types. This kind of branching process is called *damage modeling* in bivariate distribution literature. Marginally, the total number of fatal accidents follows a quasi-Poisson distribution

$$p(x) = \sum_{y=0}^{\infty} p(y, x) = \lambda_1 p (\lambda_1 p + \lambda_2 x)^{x-1} \exp[-(\lambda_1 p + \lambda_2 x)] / x!,$$

where the observation $x = 0, 1, 2, \dots, \infty$ and $(\lambda_1 p + \lambda_2 x) > 0$. The marginal mean of X is

$$E(X) = \frac{\lambda_1 p}{1 - \lambda_2},$$

and the variance is

$$\text{var}(X) = \frac{\lambda_1 p}{(1 - \lambda_2)^3}$$

In the absence of accident proneness, (that is, $\lambda_2 = 0$), the marginal mean and variance of X are $\lambda_1 p$, a unique Poisson property to be watched. Intuitively, the random observations on Y and X must be correlated and not independent. Such an intuition is backed up by the difference between marginal and conditional probability distributions. That is, $p(Y = y | X = x) \neq p(X = x | Y = y)$ and $p(X = x | Y = y) \neq p(X = x)$ where

$$\begin{aligned} p(Y = y | X = x) &= \lambda_1 [1 - p] (\lambda_1 [1 - p] + \lambda_2 [y - x])^{y-x-1} \\ &\quad \exp[-(\lambda_1 [1 - p] + \lambda_2 [y - x])] / (y - x)! \end{aligned}$$

with $y = x, x+1, \dots$. This conditional distribution is also quasi-Poisson but shifted x units on the right. It is easy to notice that $E(X | Y = y) = yp$, a regression line with slope p passing through the origin. But the reversed regression curve of Y given $X = x$ is

$$E(Y | X = x) = x + \frac{\lambda_1 [1 - p]}{[1 - \lambda_2]},$$

with an intercept equal to $\frac{\lambda_1 [1 - p]}{[1 - \lambda_2]}$

and unit slope. For random observations Y and X to be independent, a rule of thumb is $E(Y | X = x) = E(Y)$ and

Table 1 Selective Bivariate Continuous (Cumulative) Probability Distributions

Name	Probability density $f(x, y)$ or cumulative distribution function, $F(x, y) = \Pr[Y \leq y, X \leq x]$
Gaussian	$f(y, x) = [2\pi\sigma_1\sigma_2(1 - \rho^2)]^{1/2} \exp[-(y - \mu_1)^2/\sigma_1^2 - (x - \mu_2)^2/\sigma_2^2 + 2\rho(x - \mu_1)(y - \mu_2)/\sigma_1\sigma_2]$
Marshall-Olkin exponential	$F(y, x) = \exp[-\lambda_1 y - \lambda_2 x - \lambda_3 \max(y, x)]$
Bivariate logistic	$F(y, x) = [1 + \exp(-[y - \mu_1]/\sigma_1) + \exp(-[x - \mu_2]/\sigma_2) + (1 - \rho) \exp(-[y - \mu_1]/\sigma_1 - [x - \mu_2]/\sigma_2)]^{-1}$
Pareto first kind	$f(y, x) = \lambda(\lambda + 1)/\theta_1\theta_2 \left(\frac{y}{\theta_1} + \frac{x}{\theta_2} - 1 \right)^{-(\lambda+2)},$ $y > \theta_1, x > \theta_2, \lambda > 0$
Pareto second kind	$F(y, x) = 1 - \left(1 + \frac{y - \mu_1}{\theta_1} + \frac{x - \mu_2}{\theta_2} \right)^{-\lambda},$ $y > \mu_1, x > \mu_2, \lambda > 0$
Pareto third kind	$F(y, x) = 1 - 1 / \left[\left(\frac{y - \mu_1}{\theta_1} \right)^{1/\delta_1} + \left(\frac{x - \mu_2}{\theta_2} \right)^{1/\delta_2} \right],$ $y > \mu_1, x > \mu_2, \delta_1 > 0, \delta_2 > 0$
Dirichlet	$f(y, x) = \frac{\Gamma(\theta_0 + \theta_1 + \theta_2)}{\Gamma(\theta_0)\Gamma(\theta_1)\Gamma(\theta_2)} y^{\theta_1-1} x^{\theta_2-1} (1 - [y + x])^{\theta_0-1}$
Gumbel	$f(y, x) = e^{-(y+x+\theta yx)[(1+\theta y)(1+\theta x)-\theta]}$ $y, x > 0; 0 \leq \theta \leq 1$
Freund	$f(y, x) = \begin{cases} \alpha\beta' e^{-\beta'y - (\alpha+\beta-\beta')x} \\ \alpha'\beta e^{-\alpha'x - (\alpha+\beta-\alpha')y} \end{cases} \text{ if } \begin{cases} 0 \leq x \leq y \\ 0 \leq y \leq x \end{cases}$
Block-Basu	$f(y, x) = \begin{cases} \frac{\lambda\lambda_1(\lambda_2+\lambda_{12})}{\lambda_1+\lambda_2} e^{-\lambda_1 x - (\lambda_2+\lambda_{12})y} \\ \frac{\lambda\lambda_2(\lambda_1+\lambda_{12})}{\lambda_1+\lambda_2} e^{-\lambda_2 y - (\lambda_1+\lambda_{12})x} \end{cases}, \text{ if } \begin{cases} 0 < x < y < \infty \\ 0 < y < x < \infty \end{cases}$
Bivariate extreme value	$H(x, y) = \exp(-\exp(-[y - \mu_1]/\sigma_1) + \exp(-[x - \mu_2]/\sigma_2))^{-1/\rho}$

Table 2 Selective Bivariate Count Distributions

Name	Probability mass function, $f(x, y)$
Binomial	$p(y, x, n - y - x) = \binom{n}{y!x!(n - y - x)!} \theta_1^y \theta_2^x (1 - \theta_1 - \theta_2)^{n - y - x}$ $0 < \theta_1, \theta_2 < 1; y = 0, 1, 2, \dots, n; x = 0, 1, 2, \dots, n; y + x \leq n$
Inverse binomial	$p(y, x, n - y - x) = \frac{\Gamma(r + y + x)}{\Gamma(r) y! x!} \theta_1^y \theta_2^x (1 - \theta_1 - \theta_2)^r$ $0 < \theta_1, \theta_2 < 1; y = 0, 1, 2, \dots, ; x = 0, 1, 2, \dots,$
Poisson	$p(y, x) = e^{-(\theta_1 + \theta_2 + \theta_{12})} \sum_{i=0}^{\max(y, x)} \theta_1^{y-i} \theta_2^{x-i} \theta_{12}^i / (y-i)! (x-i)! i!$
Hypergeometric	$p(y, x) = \frac{\binom{Np_1}{y} \binom{Np_2}{x} \binom{N - Np_1 - Np_2}{n - y - x}}{\binom{N}{n}};$ $y = 0, 1, \dots, \min(N_1, n - x); x = 0, 1, \dots, \min(N_2, n - y);$ $\binom{a}{b} = \frac{\Gamma(a + 1)}{\Gamma(b + 1) \Gamma(a + b + 1)}$
Bivariate geometric	$p(y, x) = \binom{y + x}{y} \theta_1^y \theta_2^x (1 - \theta_1 - \theta_2)$ $y, x = 0, 1, 2, \dots; 0 < \theta_1, \theta_2 < 1$
Sarmonov-Lee family	$p(y, x) = \binom{m}{y} \binom{n}{x} \theta_1^y (1 - \theta_1)^{m-y} \theta_2^x (1 - \theta_2)^{n-x} [1 + \varpi \phi_1(y) \phi_2(x)];$ $y = 0, 1, 2, \dots, m; x = 0, 1, 2, \dots, n; 0 < \theta_1, \theta_2 < 1$
Quasi-binomial-Poisson	$p(y, x) = \lambda_1^2 p(1 - p) (\lambda_1 p + \lambda_2 x)^{x-1} (\lambda_1 [1 - p] + ([y - x] \lambda_2)^{y-x-1} \exp[-(\lambda_1 + \lambda_2 y)] / x! (y - x)!$ $y = 0, 1, 2, \dots, \infty;$ $x = 0, 1, 2, \dots, y; (\lambda_1 + \lambda_2 y) > 0;$ $\lambda_1 > 0; -1 < \lambda_2 < 1; 0 < p < 1$

$E(X|Y=y) = E(X)$. This rule is clearly not validated in this example. So the number of accidents and the number of fatal accidents must be not independent but correlated. Their correlation is

$$\rho_{Y,X} = \frac{\text{cov}(Y, X)}{\sqrt{\text{var}(Y)\text{var}(X)}} = +\sqrt{p}.$$

Also,

$$\text{var}(Y) = \frac{\lambda_1}{(1 - \lambda_2)^3},$$

$$\text{var}(X) = \frac{\lambda_1 p}{(1 - \lambda_2)^3},$$

$$\text{var}(Y|X=x) = \frac{\lambda_1[1-p]}{(1 - \lambda_2)^3},$$

and

$$\begin{aligned} \text{var}(X|Y=y) &= y^2 p[1-p] \\ &- \frac{y(y-1)\lambda_1 p(1-p)}{(\lambda_1 + \lambda_2 y)} \sum_{s=0}^{y-2} \frac{(y-2)s\lambda_2^s}{(\lambda_1 + \lambda_2 y)^s}. \end{aligned}$$

These results validate a universal result in bivariate distribution that

$$\text{var}(Y) = \text{var}(E[Y|X=x]) + E(\text{var}[Y|X=x]).$$

The correlation coefficient is a dependence measure. The importance of dependence measures in bivariate distributions cannot be overstated. Of course, all dependent measures are model based, and hence, selecting an appropriate bivariate distribution for a given set of data is vitally important. The concept of *copula* (meaning bond) eases the burden of selecting a bivariate distribution. The copula is a scale-invariant way of dealing with dependency. Only the uniform distribution over a unit square for the copula can detect the independence between bivariate. A departure from uniformity indicates the existence of dependency. What kind of dependencies can a copula detect in bivariate distributions? For a discussion of this wonderful idea, consider the cumulative distribution functions of Y , X , and their joint random variables, which are indicated respectively by $u = G(x) = \Pr[X \leq x]$, $v = F(y) = \Pr[Y \leq y]$, and $H(x, y) = \Pr[X \leq x, Y \leq y]$. The copula is then a mapping of each (x, y) in a two-dimensional domain to a unique

value $H(x, y)$ in the unit square. There is a unique copula $C(u, v)$ in the sense $H(G^{-1}[u], F^{-1}[v]) = C(u, v)$. Bivariate random variables Y and X are considered independent if and only if there is a copula validating $B(u, v) = uv$. In general, $\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v)$.

For the sake of understanding continuous bivariate distributions, consider bivariate logistic distribution (as in Table 1). Bivariate logistic distribution is employed to explain random failing of paired organs such as kidneys in diabetic patients. When their correlation coefficient $\rho = 0$, the random lifetimes Y and X are independent since

$$\begin{aligned} H(x, y) &= \exp(-[y - \mu_1]/\sigma_1) \exp(-[x - \mu_2]/\sigma_2) \\ &= F(y)G(x) \text{ and } B(u, v) = uv. \end{aligned}$$

Consider the bivariate extreme value distribution (as in Table 1) with $\sigma = 1/\rho = \sigma_2$, $\mu_1 = 0 = \mu_2$. This bivariate extreme value distribution illustrates unusual rainfall and farm damage. Its copula is $C(u, v) = \exp(-[(-1nu)^\rho + (-1nv)^\rho]^{1/\rho})$. When the parameter $\rho = 1$, the amount of rainfall and the amount of farm damage are stochastically independent.

In the case of bivariate distributions, copula is related to Spearman's correlation coefficient ρ and Kendall's rank correlation coefficient τ . Bivariate random variables Y and X are *concordant* if large values of one variate tend to be associated with large values of the other variate and smaller values of one variate are associated with small values of the other. Otherwise, the bivariate are *discordant*. Kendall's τ is simply the probability of concordance minus the probability of discordance. The copula is related to Kendall's τ as $C(u, v) = (1+\tau)/4$. For a given set of data, both ρ and τ could come out differently but should validate an inequality $-1 \leq 3\tau \leq 3\rho \leq 1$.

—Ramalingam Shanmugam

Further Reading

- Basu, A. P., & Dhar, S. K. (1995). Bivariate geometric distribution. *Journal of Applied Statistical Science*, 2, 33–44.
- Geoffroy, J. (1958, 1959). Contribution a la theorie des valeurs extremes. *Publications de l'Institut de Statistique de l'Universite de Paris*, 7, 37–121; 8, 123–184.

- Lee, M. L. T. (1966). Properties and applications of the Sarmonov family of bivariate distributions. *Communications in Statistics*, 25, 1207–1222.
- Sarmonov, O. V. (1966). Generalized normal correlation and two dimensional Frechet classes. *Soviet Doklady*, 168, 596–599.
- Shanmugam, R. (2002). A critique of dependence concepts. *Journal of Mathematical Psychology*, 46, 110–114.
- Shanmugam, R., & Singh, J. (1981). Some bivariate probability models applicable to traffic accidents and fatalities. In C. Tallie, G. P. Patil, & B. A. Baldessari (Eds.), *Statistical distributions in scientific work: Vol. 6* (pp. 95–103). Hingham, MA: Reidel.

BONFERRONI, CARLO EMILIO (1892–1960)

Bonferroni was born in Bergamo (near Milan), Italy, on January 28, 1892. Educated at Turin University, his first post was assistant professor in financial mathematics, mechanics, and geometry at the Turin Polytechnic. In 1923, he was appointed professor of financial mathematics at the Economics Institute in Bari, where he served 7 years as rector.

In 1933 he moved to Firenze (Florence), where he remained until his death. During his time in Firenze, he filled a variety of administrative posts. For example, in the immediate postwar years, he acted both as head of the statistics faculty at Bocconi University, Milan, and as head of the Faculty of Architecture in Florence.

His work on inequalities, published in 1935 and 1936, represented only a small part of his interests. For example, his inaugural lecture was concerned with the foundations of probability (which he viewed as the limit of relative frequency when the entire population is sampled).

The handwritten notes he produced for his students reveal his deep insights into mathematics—often revealed through neat and idiosyncratic solutions. By all accounts, he was a sensitive and kind-hearted man and a gentleman. He was a talented pianist and also composed music. In his younger days, he was a keen glacier walker. His garden was described as enchanting.

Bonferroni died on August 18, 1960, in Firenze.

—Graham Upton

Further Reading

Galambos, J., & Simonelli, I. (1996). *Bonferroni-type inequalities with applications*. New York: Springer.

Carlo Emilio Bonferroni biography and readings: <http://www.aghmed.fsnet.co.uk/bonf/bonf.html>

BONFERRONI TEST

The more tests we perform on a set of data, the more likely we are to reject the null hypothesis when it is true (a Type I error). This is a consequence of the logic of hypothesis testing: We reject the null hypothesis if we witness a rare event. But the larger the number of tests, the easier it is to find rare events, and therefore, the easier it is to make the mistake of thinking that there is an effect when there is none. This problem is called the *inflation* of the alpha level. One strategy for preventing it is to correct the alpha level when performing multiple tests. Making the alpha level more stringent (i.e., smaller) will create fewer errors, but it may also make real effects harder to detect.

The Different Meanings of Alpha

Maybe researchers perform more and more statistical tests on one set of data because computers make statistical analyses easy to run. For example, brain imaging researchers will routinely run millions of tests to analyze an experiment. Running so many tests increases the risk of false alarms. To illustrate, imagine the following “pseudoexperiment”:

I toss 20 coins, and I try to force the coins to fall heads up. I know that, from the binomial test, the null hypothesis is rejected at the $\alpha = .05$ level if the number of heads is greater than 14. I repeat this experiment 10 times.

Suppose that one trial gives the “significant” result of 16 heads versus 4 tails. Did I influence the coins on that occasion? Of course not, because the larger the number of experiments, the greater the probability of encountering a low-probability event (like 16 versus 4). In fact, waiting long enough is a sure way of detecting rare events!

Probability in the Family

A *family of tests* is the technical term for a series of tests performed on a set of data. In this section, we show how to compute the probability of rejecting the null hypothesis at least once in a family of tests when the null hypothesis is true.

For convenience, suppose that we set the significance level at $\alpha = .05$. For each test (i.e., one trial in the example of the coins), the probability of making a Type I error is equal to $\alpha = .05$. The events “making a Type I error” and “not making a Type I error” are *complementary events* (they cannot occur simultaneously). Therefore the probability of not making a Type I error on one trial is equal to

$$1 - \alpha = 1 - .05 = .95.$$

Recall that when two events are independent, the probability of observing these two events together is the product of their probabilities. Thus, if the tests are independent, the probability of not making a Type I error on the first *and* the second tests is

$$.95 \times .95 = (1 - .05)^2 = (1 - \alpha)^2.$$

With three tests, we find that the probability of not making a Type I error on all tests is

$$.95 \times .95 \times .95 = (1 - .05)^3 = (1 - \alpha)^3.$$

For a family of C tests, the probability of not making a Type I error for the *whole family* is

$$(1 - \alpha)^C.$$

For our example, the probability of not making a Type I error on the family is

$$(1 - \alpha)^C = (1 - .05)^{10} = .599.$$

Now, what we are looking for is the probability of making one or more Type I errors on the family of tests. This event is the complement of the event of *not making a Type I error on the family*, and therefore it is equal to

$$1 - (1 - \alpha)^C.$$

For our example, we find

$$1 - (1 - .05)^{10} = .401.$$

So, with an α level of .05 for *each* of the tests, the probability of wrongly rejecting the null hypothesis is .401.

This example makes clear the need to distinguish between two meanings of α when performing multiple tests:

- The probability of making a Type I error when dealing only with a specific test. This probability is denoted $\alpha[PT]$ (pronounced “alpha per test”). It is also called the *testwise* alpha.
- The probability of making at least one Type I error for the whole family of tests. This probability is denoted $\alpha[PF]$ (pronounced “alpha per family of tests”). It is also called the *familywise* or the *experimentwise* alpha.

A Monte Carlo Illustration

A Monte Carlo simulation can illustrate the difference between $\alpha[PT]$ and $\alpha[PF]$. The Monte Carlo technique consists of running a simulated experiment many times using random data. This gives the pattern of results that happens on the basis of chance.

Here six groups with 100 observations per group were created with data randomly sampled from the same normal population. By construction, H_0 is true (i.e., all population means are equal). Call that procedure an experiment. We performed five independent tests from these six groups. For each test, we computed an F test. If its probability was smaller than $\alpha = .05$, the test was declared significant (i.e., $\alpha[PT]$ is used). We performed this experiment 10,000 times. Therefore, there were 10,000 experiments, 10,000 families, and $5 \times 10,000 = 50,000$ tests. The results of this simulation are given in Table 1.

Table 1 shows that H_0 is rejected for 2,403 tests of more than 50,000 tests performed. From these data, an estimation of $\alpha[PT]$ is computed as

Table 1 Results of a Monte Carlo Simulation:
Numbers of Type I Errors When
Performing $C = 5$ Tests for 10,000 Families
When H_0 Is True*

Number of Families With X Type I Errors	X : Number of Type I Errors per Family	Number of Type I Errors
7,868	0	0
1,907	1	1,907
192	2	384
20	3	60
13	4	52
0	5	0
10,000		2,403

Note: *For example, 192 families out of 10,000 have two Type I errors; this gives $2 \times 192 = 384$ Type I errors.

$$\alpha[PT] = \frac{\text{number of significant tests}}{\text{total}} \quad (1)$$

$$\text{number of tests} = \frac{2,403}{50,000} = .0479.$$

This value falls close to the theoretical value of $\alpha = .05$.

For 7,868 families, no test reaches significance. Equivalently, for 2,132 families ($10,000 - 7,868$), at least one Type I error is made. From these data, $\alpha[PF]$ can be estimated as

$$\alpha[PF] = \frac{\text{number of families with at least 1 Type I error}}{\text{total number of families}} \quad (2)$$

$$= \frac{2,132}{10,000} = .2132.$$

This value falls close to the theoretical value of

$$\alpha[PF] = 1 - (1 - \alpha[PT])^C = 1 - (1 - .05)^5 = .226$$

How to Correct for Multiple Tests: Šidàk, Bonferroni, Boole, Dunn

Recall that the probability of making at least one Type I error for a family of C tests is

$$\alpha[PF] = 1 - (1 - \alpha[PT])^C$$

This equation can be rewritten as

$$\alpha[PF] = 1 - (1 - \alpha[PT])^{1/C}$$

This formula—derived assuming independence of the tests—is sometimes called the Šidàk equation. It shows that in order to reach a given $\alpha[PF]$ level, we need to adapt the $\alpha[PT]$ values used for each test.

Because the Šidàk equation involves a fractional power, it is difficult to compute by hand, and therefore several authors derived a simpler approximation, which is known as the Bonferroni (the most popular name), or Boole, or even Dunn approximation. Technically, it is the first (linear) term of a Taylor expansion of the Šidàk equation. This approximation gives

$$\alpha[PT] \approx \frac{\alpha[PF]}{C}.$$

Šidàk and Bonferroni are linked to each other by the inequality

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} \geq \frac{\alpha[PF]}{C}.$$

They are, in general, very close to each other, but the Bonferroni approximation is pessimistic (it always does worse than the Šidàk equation). Probably because it is easier to compute, the Bonferroni approximation is better known (and cited more often) than the exact Šidàk equation.

The Šidàk-Bonferroni equations can be used to find the value of $\alpha[PT]$ when $\alpha[PF]$ is fixed. For example, suppose that you want to perform four independent tests, and because you want to limit the risk of making at least one Type I error to an overall value of $\alpha[PF] = .05$, you will consider a test significant if its associated probability is smaller than

$$\alpha[PF] = 1 - (1 - \alpha[PT])^{1/C} = 1 - (1 - .05)^{1/4} = .0127.$$

With the Bonferroni approximation, a test reaches significance if its associated probability is smaller than

$$\alpha[PT] = \frac{\alpha[PF]}{C} = \frac{.05}{4} = .0125,$$

which is very close to the exact value of .0127.

Correction for Nonindependent Tests

The Šidàk equation is derived assuming independence of the tests. When they are not independent, it gives a lower bound, and then

$$\alpha[PF] \leq 1 - (1 - \alpha[PT])^C.$$

As previously, we can use a *Bonferroni* approximation because

$$\alpha[PF] < C\alpha[PT].$$

Šidàk and Bonferroni are related by the inequality

$$\alpha[PF] \leq 1 - (1 - \alpha[PT])^C < C\alpha[PT].$$

The Šidàk and Bonferroni inequalities can also be used to find a correction on $\alpha[PT]$ in order to keep $\alpha[PF]$ fixed. The Šidàk inequality gives

$$\alpha[PT] \approx 1 - (1 - \alpha[PF])^{1/C}$$

This is a conservative approximation because the following inequality holds:

$$\alpha[PT] \geq 1 - (1 - \alpha[PF])^{1/C}$$

The Bonferroni approximation gives

$$\alpha[PT] \approx \frac{\alpha[PF]}{C}.$$

Splitting Up $\alpha[PF]$ With Unequal Slices

With the Bonferroni approximation, we can make an unequal allocation of $\alpha[PF]$. This works because with the Bonferroni approximation, $\alpha[PF]$ is the sum of the individual $\alpha[PT]$:

$$\alpha[PF] \approx C\alpha[PT] = \underbrace{\alpha[PT] + \alpha[PT] + \dots + \alpha[PT]}_{C \text{ times}}.$$

If some tests are judged more important a priori than some others, it is possible to allocate $\alpha[PF]$ unequally. For example, suppose we have three tests that we want to test with an overall $\alpha[PF] = .05$, and we think that the first test is the most important of the set. Then we can decide to test it with $\alpha[PT] = .04$ and share the remaining value, $.01 = .05 - .04$, between the last two tests, which will be evaluated each with a value of $\alpha[PT] = .005$. The overall Type I error for the family is equal to $\alpha[PF] = .04 + .005 + .005 = .05$, which was indeed the value we set beforehand. It should be emphasized, however, that the (subjective) importance of the tests and the unequal allocation of the individual $\alpha[PT]$ should be decided a priori for this approach to be statistically valid. An unequal allocation of the $\alpha[PT]$ can also be achieved using the Šidàk inequality, but it is more computationally involved.

Alternatives to Bonferroni

The Šidàk-Bonferroni approach becomes very conservative when the number of comparisons becomes large and when the tests are not independent (e.g., as in brain imaging). Recently, some alternative approaches have been proposed to make the correction less stringent. A more recent approach redefines the problem by replacing the notion of $\alpha[PF]$ with the false discovery rate, which is defined as the ratio of the number of Type I errors to the number of significant tests.

—Hervé Abdi

Further Reading

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.
- Games, P. A. (1977). An improved *t* table for simultaneous control on *g* contrasts. *Journal of the American Statistical Association*, 72, 531–534.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–803.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons*. New York: Cambridge University Press.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–584.
- Šidák, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626–633.

BOWKER PROCEDURE

It is often of interest to examine changes in the categorical responses taken from participants before and then after some treatment condition is imposed (i.e., to evaluate repeated measurements of the same participants, using them as their own controls). In 1947, the psychologist Quinn McNemar developed a simple procedure for comparing differences between the proportions in the before and after responses for two categories. In 1948, the statistician Albert Bowker expanded on McNemar's work and developed a test for symmetry that evaluates the changes in before and after responses in contingency tables when there are multiple categories.

Bowker's procedure has been used broadly both in the social and behavioral sciences and in medical research, and some attention has been given to applications in advertising, public relations, and marketing research, wherein it may be desirable to evaluate the significance of changes in attitudes, opinions, and beliefs.

Development

The responses from a sample of n' individuals over two periods of time may be tallied into an $r \times c$ table (where r , the number of rows, equals c , the number of columns) of cross-classifications, as shown in Table 1.

With respect to the population from which the aforementioned sample was taken, let p_{ij} be the probability of responses to the i th category before the treatment condition was imposed and the j th category after. The marginal probabilities before and after treatment sum to unity. That is, $p_{1.} + p_{2.} + \dots + p_{r.} = 1$ and $p_{.1} + p_{.2} + \dots + p_{.c} = 1$.

Testing for Significance of Changes in Related Proportions

In order to investigate changes in repeated measurements, the null hypothesis is that of symmetry:

$$H_0: p_{ij} = p_{ji} \text{ for all } i > j.$$

The alternative is that at least one pair of symmetric probabilities is unequal:

$$H_1: p_{ij} \neq p_{ji} \text{ for any } i > j.$$

That is, the null hypothesis tested is conditioned on those

$$n = \sum_{\forall i \neq j} x_{ij}$$

individuals whose responses change, where the probability (p_{ij}) of a switch from response i to response j is equal to the probability (p_{ji}) of a switch from response j to response i , and this probability is 0.5.

The Bowker test statistic B , written as

$$B = \sum_{i=j+1}^r \sum_{j=1}^c \frac{(x_{ij} - x_{ji})^2}{x_{ij} + x_{ji}},$$

has a chi-square distribution with u degrees of freedom where $u = r(r-1)/2 = c(c-1)/2$ since $r = c$. The

Table 1 $r \times c$ Table of Cross-Classifications for a Sample of n' Subjects

$I \backslash II$					Totals
	1	2	...	$c = r$	
1	x_{11}	x_{12}	...	x_{1c}	$x_{1.}$
2	x_{21}	x_{22}	...	x_{2c}	$x_{2.}$
...	...	...	...	...	...
$r = c$	x_{r1}	x_{r2}	...	x_{rc}	$x_{r.}$
Totals	$x_{.1}$	$x_{.2}$	...	$x_{.c}$	n'

Notes: I = Time period I (before treatment) in a repeated measurements experiment; II = Time period II (after treatment) in a repeated measurements experiment; r = number of rows (number of categories); c = number of columns (number of categories); n' = sample size.

null hypothesis can be rejected at the α level of significance if

$$B > \chi_{v,1-\alpha}^2.$$

A Posteriori Comparisons

If the null hypothesis is rejected, the researcher Alan Stuart suggested a multiple comparison procedure that permits the development of a post hoc evaluation of changes in the correlated proportions (i.e., marginal probabilities) for each response category versus the $c - 1$ other categories combined. Thus, regardless of the size of the initial $c \times c$ table of cross-classifications, this process allows for the formation of a set of $c \times 2$ tables, one for each of the c categories versus all the other $c - 1$ categories combined. These 2×2 tables take the form

<i>After</i> <i>Before</i>	<i>i</i>	<i>Not i</i>	<i>Totals</i>
<i>i</i>	x_{ii}	$x_{ii'}$	$x_{i.}$
<i>Not i</i>	$x_{i'i}$	$x_{i'i'}$	$x_{i'.$
<i>Totals</i>	$x_{.i}$	$x_{.i'}$	n'

for all $i = 1, \dots, c$.

The critical ranges for each of these c a posteriori comparisons arise from the standard error of the differences in two related proportions as used in McNemar-type confidence intervals.

With an experimentwise error rate α , each of the possible c pairwise comparisons is made, and the decision rule is to declare column classification i different from row classification i if

$$|\hat{p}_{.i} - \hat{p}_{i.}| > \sqrt{\chi_{\alpha, (c-1)}^2} \cdot \sqrt{\frac{\hat{p}_{.i} \hat{p}_{.i'}}{n'} + \frac{\hat{p}_{i.} \hat{p}_{i'}}{n'} - \frac{2(\hat{p}_{ii} - \hat{p}_{.i} \hat{p}_{i.})}{n'}}$$

where $\hat{p}_{.i} = \frac{x_{.i}}{n'}$ and $\hat{p}_{i.} = \frac{x_{i.}}{n'}$ and

where i' is the complement of i , the combined responses from all $c - 1$ other classifications. That is,

the classification in column i and classification in row i are declared significantly different if $|\hat{p}_{.i} - \hat{p}_{i.}|$, the absolute difference in the sample proportions of “success” before and after a treatment intervention, exceeds a critical range given by the product

$$\text{of } \sqrt{\chi_{\alpha, (c-1)}^2} \text{ and } \sqrt{\frac{\hat{p}_{.i} \hat{p}_{.i'}}{n'} + \frac{\hat{p}_{i.} \hat{p}_{i'}}{n'} - \frac{2(\hat{p}_{ii} - \hat{p}_{.i} \hat{p}_{i.})}{n'}}.$$

Applying the Bowker Procedure

Consider the following hypothetical example: Suppose a consumer panel consists of $n' = 150$ individuals who have stated their preference for Toyota Camry, Honda Accord, and Nissan Maxima vehicles. The participants were each asked which of these vehicles they were most likely to choose for their next car purchase. Of the 150 participants, 45 stated they were most likely to purchase a Toyota Camry, 66 were most likely to purchase a Honda Accord, and 39 were most likely to purchase a Nissan Maxima.

Suppose the panelists are then presented with a consumer-based rating of customer satisfaction, product quality, and buyer behavior for these vehicles, such as would be found in a J. D. Power automobile report. The ratings provide detailed information about these three vehicles and rank the vehicles overall from best to worst as follows: Toyota Camry, Nissan Maxima, and Honda Accord.

Following this exposure to the research literature, the individuals are asked once again to answer the above question. Of the 45 panelists who initially stated they were most likely to purchase a Toyota Camry, 40 remained consistent in their response, 3 stated they were now more likely to purchase a Honda Accord, and 2 were now more likely to purchase a Nissan Maxima. Of the 66 panelists who initially stated they were most likely to purchase a Honda Accord, 41 remained consistent in their response, 14 stated they were now more likely to purchase a Toyota Camry, and 11 were now more likely to purchase a

Nissan Maxima. Of the 39 panelists who initially stated they were most likely to purchase a Nissan Maxima, 29 remained consistent in their response, 6 stated they were now more likely to purchase a Toyota Camry, and 4 were now more likely to purchase a Honda Accord.

The results are displayed in Table 2.

Bowker's test seeks to measure the ability of the automobile ratings to affect consumer preferences. The null hypothesis for the test of symmetry is

$$H_0: p_{ij} = p_{ji} \text{ for all } i > j.$$

That is, a panelist's car preference is not affected by the information in the automobile ratings. For example, having read the report, a panelist who initially was leaning toward purchasing a Toyota Camry is now just as likely to feel more inclined to purchase a Nissan Maxima as a panelist would be to switch from most likely to purchase a Nissan Maxima to more likely to purchase a Toyota Camry. This null hypothesis of symmetry may be tested against the alternative:

$$H_1: p_{ij} \neq p_{ji} \text{ for any } i > j.$$

That is, exposure to the automobile ratings does influence one's preference for a particular vehicle.

For these data, the Bowker test for symmetry enables an exact test of the null hypothesis with the chi-square probability distribution, where $r = c = 3$ and $v = 3(3 - 1)/2 = 3$ degrees of freedom with a "stated" level of significance α . The Bowker test statistic B is calculated as follows, based on the response

tallies in symmetric positions off the main diagonal of the 3×3 contingency table (i.e., cells x_{12} versus x_{21} , x_{13} versus x_{31} , and x_{23} versus x_{32} from Table 2:

$$B = \frac{(14 - 3)^2}{17} + \frac{(6 - 2)^2}{8} + \frac{(4 - 11)^2}{15} \\ = 12.384.$$

Since $B > \chi_{0.05,3}^2 = 7.815$, the null hypothesis is rejected at the $\alpha = 0.05$ level of significance. The p value is 0.0062. Thus, exposure to the automobile ratings does influence one's preference for a particular vehicle. At least one of the three pairs of symmetric probabilities off the main diagonal of the 3×3 contingency table is unequal, or preference for at least one of the three automobiles has changed significantly.

Given that the null hypothesis is rejected, to determine which of the three automobiles displayed significant change in preferences as a result of the automobile ratings, post hoc evaluations for the $c = 3$ pairwise differences in related proportions (i.e., marginal probabilities) are made. The data, collapsed into three 2×2 contingency tables, are presented in Table 3.

The critical ranges for these pairwise comparisons of preferences for the particular automobiles before and then after the automobile ratings were examined are obtained in Table 4, and the pairwise comparisons of differences in proportions of preference are evaluated in Table 5.

From Table 5 it is clear that preference for the Toyota Camry significantly increased as a result of the automobile ratings while preference for the Honda Accord significantly decreased. However, the observed decline in preference for the Nissan Maxima is not statistically significant.

Table 2 Hypothetical Results of a Marketing Campaign

Before \ After				
	Toyota	Honda	Nissan	Totals
Toyota	$X_{11} = 40$	$x_{12} = 3$	$x_{13} = 2$	$x_{1\cdot} = 45$
Honda	$X_{21} = 14$	$x_{22} = 41$	$x_{23} = 11$	$x_{2\cdot} = 66$
Nissan	$X_{31} = 6$	$x_{32} = 4$	$x_{33} = 29$	$x_{3\cdot} = 39$
Totals	$X_{\cdot 1} = 60$	$x_{\cdot 2} = 48$	$x_{\cdot 3} = 42$	$n' = 150$

Discussion

For the special case where $r = c = 2$, the Bowker statistic is identical to the McNemar test for the significance of changes. Unlike McNemar's procedure for a 2×2 contingency table, Bowker's test of the null hypothesis of symmetry in a $c \times c$ contingency table is not equivalent to a test of the null hypothesis of equality of correlated proportions (i.e., marginal probability

Table 3 A Posteriori Analysis: Collapsing for Pairwise Comparisons

<i>Before \ After</i>	<i>Toyota</i>	<i>Not Toyota</i>	<i>Totals</i>
Toyota	$X_{11} = 40$	$x_{12} = 5$	$x_{1.} = 45$
Not Toyota	$X_{21} = 20$	$x_{22} = 85$	$x_{2.} = 105$
Totals	$x_{.1} = 60$	$x_{.2} = 90$	$n' = 150$

<i>Before \ After</i>	<i>Honda</i>	<i>Not Honda</i>	<i>Totals</i>
Honda	$X_{11} = 41$	$x_{12} = 25$	$x_{1.} = 66$
Not Honda	$X_{21} = 7$	$x_{22} = 77$	$x_{2.} = 84$
Totals	$x_{.1} = 48$	$x_{.2} = 102$	$n' = 150$

<i>Before \ After</i>	<i>Nissan</i>	<i>Not Nissan</i>	<i>Totals</i>
Nissan	$X_{11} = 29$	$x_{12} = 10$	$x_{1.} = 39$
Not Nissan	$X_{21} = 13$	$x_{22} = 98$	$x_{2.} = 111$
Totals	$x_{.1} = 42$	$x_{.2} = 108$	$n' = 150$

distributions). In 1955, Stuart proposed a test for the latter and, on rejection of the null hypothesis, proposed a set of McNemar-type simultaneous confidence intervals for a posteriori evaluation of differences in the c pairs of correlated proportions.

Taking advantage of the fact that testing a null hypothesis of symmetry is equivalent to testing a null hypothesis of correlated proportions in the 2×2 contingency table, the a posteriori McNemar-type simultaneous confidence intervals proposed by Stuart can be adapted to the Bowker procedure by collapsing the $c \times c$ contingency table on the main diagonal into a set of c 2×2 contingency tables so that one may test globally, for each category, for the significance of change in the proportion of respondents switching from one category to all others combined versus switching to a category from any other. In the hypothetical example presented here, the intervention through exposure to the automobile ratings caused a significant shift in preference to the Toyota Camry and a significant shift away from the Honda Accord.

As a result of the treatment intervention, if the gains are transitive, a rejection of the null hypothesis

Table 4 Computation of Critical Ranges for Post Hoc Pairwise Comparisons

Automobiles	$\sqrt{\chi^2_{\alpha, (c-1)}} \cdot \sqrt{\frac{\hat{p}_{.i} \hat{p}_{.i'}}{n'} + \frac{\hat{p}_{i.} \hat{p}_{i'}}{n'} - \frac{2(\hat{p}_{ii} - \hat{p}_{.i} \hat{p}_{i.})}{n'}}$
Toyota vs. not Toyota	$\sqrt{5.991} \cdot \sqrt{\frac{(0.40)(0.60)}{150} + \frac{(0.30)(0.70)}{150} - \frac{2[(0.2667) - (0.40)(0.30)]}{150}} = 0.0791$
Honda vs. not Honda	$\sqrt{5.991} \cdot \sqrt{\frac{(0.32)(0.68)}{150} + \frac{(0.44)(0.56)}{150} - \frac{2[(0.2733) - (0.32)(0.44)]}{150}} = 0.0892$
Nissan vs. not Nissan	$\sqrt{5.991} \cdot \sqrt{\frac{(0.28)(0.72)}{150} + \frac{(0.26)(0.74)}{150} - \frac{2[(0.1933) - (0.28)(0.26)]}{150}} = 0.0782$

Table 5 Post Hoc Pairwise Comparisons of Changes in Perceptions for the Automobiles

Automobiles	$ \hat{p}_{.i} - \hat{p}_{i.} $	Critical range	Decision rule
Toyota vs. not Toyota	$ \hat{p}_{.1} - \hat{p}_{1.} = 0.1000$	0.0791	Significant
Honda vs. not Honda	$ \hat{p}_{.2} - \hat{p}_{2.} = 0.1200$	0.0892	Significant
Nissan vs. not Nissan	$ \hat{p}_{.3} - \hat{p}_{3.} = 0.0200$	0.0782	Not Significant

by the Bowker test should lead to global findings, as shown here.

In recent years, the biostatisticians Warren L. May and William D. Johnson have thoroughly researched the issue of symmetry among several proportions and proposed an alternative approach to the Bowker procedure, along with simultaneous confidence intervals for a posteriori analysis.

Conclusions

It is essential to a good data analysis that the appropriate statistical procedure be applied to a specific situation. The Bowker test may be used when studying symmetry among several proportions based on related samples. A researcher unaware of the procedure may employ an inappropriate chi-square test for the $c \times c$ contingency table and draw incorrect conclusions.

The Bowker test is quick and easy to perform. The only assumption is that the before and after responses of each participant are categorized into a $c \times c$ table.

The pedagogical advantage of the a posteriori multiple comparisons based on McNemar-type confidence intervals is that they demonstrate that all n' participants are being evaluated. The initial B test statistic itself is conditioned on a reduced set of participants, the “brand switching” panelists off the main diagonal of the cross-classification table.

—Nicole B. Koppel and Mark L. Berenson

Further Reading

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572–574.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences* (167–171). Monterey, CA: Brooks/Cole.

May, W. L., & Johnson, W. D. (1997). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine*, 16, 2127–2136.

May, W. L., & Johnson, W. D. (2001). Symmetry in square contingency tables: Tests of hypotheses and confidence interval construction. *Journal of Biopharmaceutical Statistics*, 11(1–2), 23–33.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157.

Stuart, A. (1955). A test of homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412–416.

Stuart, A. (1957). The comparison of frequencies in matched samples. *British Journal of Statistical Psychology*, 10, 29–32.

BOX PLOT (BOX AND WHISKER PLOT)

The box and whisker plot was developed by John Tukey to summarize visually the important characteristics of a distribution of scores. The five descriptive statistics included on a box plot are the minimum and the maximum scores (i.e., the extremes of the distribution), the median (i.e., the middle score), and the 1st (Q_1) and 3rd (Q_3) quartiles. Together these statistics are useful in visually summarizing, understanding, and comparing many types of distributions.

In a box plot, the crossbar indicates the median, and the length (i.e., height) of the box indicates the interquartile range (IQR) (i.e., the central 50% of the data values). The length of the whiskers indicates the range of scores that are included within 1.5 IQRs below and above the 1st and 3rd quartiles, respectively.

Box plots are particularly useful for investigating the symmetry of a distribution and for detecting inconsistent values and outliers. Outliers, which are scores that are more than 1.5 IQRs below Q_1 or above Q_3 , are plotted individually on a box plot. In a normal distribution, about 1% of the scores will fall outside the box and whiskers. The symmetry of a distribution

is indicated by where the median bifurcates the box (in a symmetrical distribution, the median is close to the center of the box) and by the length of the whiskers (in a distribution with symmetrical tails, the whiskers are of similar length).

Figure 1 summarizes the descriptive statistics and displays the box plots for 100 randomly selected IQ scores and for the subset of all scores that are greater than 99. Figure 1 shows that variable “Random” is roughly symmetrical, with three low IQ-score outliers. Variable “>99” is slightly positively skewed (the median is closer to Q_1 than to Q_3 , the upper whisker is longer than the lower whisker, and there are no outliers).

The box plot is a classic exploratory data analysis tool that is easy to construct and interpret. It is resistant to small changes in the data (up to 25% of the scores can change with little effect on the plot) because its major components are the median and the quartiles. When one is interpreting a box plot, the following limitations should be noted:

1. Quartiles (also called “hinges”) are defined differently in various computer programs, and these differences can produce very different-looking plots when sample sizes are small.
2. Although the 1.5 IQR value is used in most computer programs to draw the whiskers and define outliers, this value is not universal.
3. Using the box plot to detect outliers is a conservative procedure. It identifies an excessive number of outlying values.
4. Box plots may not have asymmetrical whiskers when there are gaps in the data.
5. Because the length of the box indicates only the spread of the distribution, multimodality and other fine features in the center of the distribution are not conveyed readily by the plot.

In order to address some of these limitations, other forms of the box plot have been developed. For example, the *variable-width box plot* is used to indicate relative sample size, the *notched box plot* is used to indicate the confidence interval of the median to

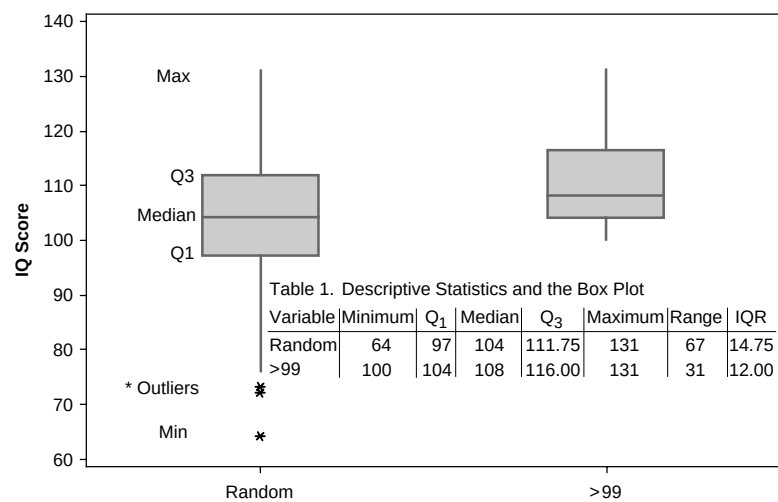


Figure 1 Box Plots of 100 Normally Distributed IQ Scores and a Subset of All IQ Scores Greater Than 99

enable comparisons between centers of distributions, and the *violin plot* combines the box plot with density traces to show multimodality and other fine-grain features of the distribution.

—Ward Rodriguez

See also Exploratory Data Analysis; Median; Range

Further Reading

- Benjamini, Y. (1988). Opening the box of a boxplot. *American Statistician*, 42, 257–262.
- Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.
- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, 43, 50–54.
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician*, 52, 181–184.
- McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *American Statistician*, 32, 12–16.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Cefai, C. (2004). Pupil resilience in the classroom: A teacher’s framework. *Emotional & Behavioural Difficulties*, 9(3), 149–170.

This article describes the development of a teacher's framework for identifying a number of primary school classes in Malta characterized by high levels of pupil resilience, namely socioemotional competence and educational engagement. The article defines resilience as a proactive, contextual, and relational phenomenon concerning all pupils, irrespective of individual characteristics or background. The author, from the University of Malta, outlines and discusses the construction, administration, and scoring of a seven-item framework, followed by an analysis of responses from 22 teachers who rated 465 pupils in their classes, on the basis of which three classes in each school were selected for further study. **Box plots** are used to present the frequency of behavior and the level of variability for both total resilience and individual component scores. The conclusion suggests how schools and teachers may use the framework as a descriptive tool in their efforts to promote socioemotional and cognitive competence.

BRACKEN BASIC CONCEPT SCALE—REVISED

The Bracken Basic Concept Scale—Revised (BBCS-R) is an instrument designed to assess the basic concept development of children in the age range of 2 years 6 months through 7 years 11 months. The BBCS-R measures children's comprehension of 308 foundational and functionally relevant educational concepts in 11 subtests or concept categories. Of the 11 subtests on the BBCS-R, the first 6 are combined into one score and compose the BBCS-R School Readiness Composite (SRC). The SRC can be used to assess children's knowledge of those readiness concepts that parents and preschool and kindergarten teachers traditionally teach children in preparation for their formal educations (e.g., colors, shapes, sizes). The remaining subtests (7–11) each produce separate scaled scores and assess important concepts that parents and teachers often fail to teach in any systematic manner (e.g., textures/materials, time/sequence). These latter subtests and the SRC

Table 1 BBCS-R Subtests, Composites, and Average Coefficient Alpha Across Age Levels

<i>Subtests/Composites</i>	<i>Alpha</i>
1. Colors	*
2. Letters	*
3. Numbers/counting	*
4. Sizes	*
5. Comparisons	*
6. Shapes	*
School Readiness Composite	.91
7. Direction/position	.97
8. Self-/social awareness	.93
9. Texture/material	.93
10. Quantity	.95
11. Time/sequence	.93
Total test score	.98

Note: *The first six subtests do not produce individual scores but contribute to the School Readiness Composite; hence, only School Readiness Composite reliability was calculated and presented.

combine and form the BBCS-R Total Test Score. Table 1 presents the BBCS subtests, the test composites, and average cross-sample coefficient alpha estimates of internal consistency.

Historically, concepts have been identified as fundamental agents or building blocks of intelligence, and acquisition of basic concepts has been shown to be strongly related to children's overall intellectual development. Importantly, the BBCS-R is a developmentally sensitive measure of cognitive and linguistic concept attainment across cultures. Thus, the concepts assessed on the BBCS-R are more universal and fundamental in importance than are the graded vocabulary words found typically on measures of receptive vocabulary that are not conceptually oriented. Ironically, researchers have demonstrated that the administration directions of many preschool and primary tests of intelligence also are replete with basic

concepts, which may render intelligence test directions the first conceptual challenge children face on tests of intelligence.

Although intelligence is not a construct that is easily improved or sustained through remedial efforts, basic concepts can be directly targeted for instruction, and students can achieve significant and sustainable growth in concept acquisition when direct instruction is linked to BBCS-R assessment results. The *Bracken Concept Development Program* was developed to create a direct assessment–instruction linkage with the BBCS-R.

—Bruce A. Bracken

Further Reading

- Bracken, B. A. (1986). Incidence of basic concepts in five commonly used American tests of intelligence. *School Psychology International*, 7, 1–10.
- Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised*. San Antonio, TX: Psychological Corporation.
- Bracken, B. A., Barona, A., Bauermeister, J. J., Howell, K. K., Poggioli, L., & Puente, A. (1990). Multinational validation of the Bracken Basic Concept Scale. *Journal of School Psychology*, 28, 325–341.
- Bracken, B. A., Howell, K. K., Harrison, T. E., Stanford, L. D., & Zahn, B. H. (1991). Ipsative subtest pattern stability of the Bracken Basic Concept Scale and the Kaufman Assessment Battery for Children in a preschool sample. *School Psychology Review*, 20, 309–324.
- Howell, K. K., & Bracken, B. A. (1992). Clinical utility of the Bracken Basic Concept Scale as a preschool intellectual screener: Comparison with the Stanford-Binet for Black children. *Journal of Clinical Child Psychology*, 21, 255–261.
- Kaufman, A. S. (1978). The importance of basic concepts in the individual assessment of preschool children. *Journal of School Psychology*, 16, 208–211.
- Laughlin, T. (1995). The school readiness composite of the Bracken Basic Concept Scale as an intellectual screening instrument. *Journal of Psychoeducational Assessment*, 13, 294–302.
- Wilson, P. (2004). A preliminary investigation of an early intervention program: Examining the intervention effectiveness of the Bracken Concept Development Program and the Bracken Basic Concept Scale—Revised with Head Start students. *Psychology in the Schools*, 41, 301–311.

BRUNO, JAMES EDWARD (1940–)

James Edward Bruno was born to Mr. and Mrs. John Bruno in Brooklyn, New York, on December 12, 1940, as part of a set of twins.

Jim grew up in Brooklyn and Long Island before his parents relocated to Pomona, California. Young Jim was always an excellent athlete and scholar. After high school, he attended the University of California, Los Angeles (UCLA). Jim's devotion to UCLA was so strong that he would complete his B.S., M.S., and Ph.D. degrees there. An outstanding educational researcher, he was hired by UCLA as a faculty member in educational policy.

Bruno's early publications applied a balanced blend of engineering, statistics, and economics to education problems. He successfully demonstrated the use of Monte Carlo methods and linear programming in helping school districts develop their substitute teachers pool and states develop school finance programs. His mathematical skills and astute observations of human behavior led him to develop a mathematical model of voter preferences that helped John Tunney get elected to the U.S. Senate. While at the RAND Corporation, Bruno met Dr. Emir Shuford and became interested in Shuford's research on admissible probability measurement and two-dimensional, confidence weighted testing. With his skills as a scientific programmer, Bruno built the first working programs for scoring and reporting tests using a modification of this system and based on an optical scanning format. Bruno would later call this procedure for assessment *information reference testing*. Organizations such as the Federal Aviation Administration and the North Atlantic Treaty Organization used this method to examine the information quality of workers who held important and critical positions. Bruno received a patent for this system of information quality assessment. Knowledge Factor, of Denver, Colorado, is applying the technology to areas of corporate training.

Later Bruno's research turned to the measurement, perception, and differential use of time and its impact

on human behavior and demonstrated that at-risk children had an entirely different concept of time with regard to past, present, and future than those children not at risk and that these differences result in antisocial behaviors. Bruno's latest research interest involves the systematic examination of policy issues as they relate to negative geographical space, social well-being, and access to equal education opportunity. This research involves the ways geographical space shapes adolescents' behavior and impacts their attitudes toward schooling and society.

Bruno has nearly 200 publications. He is married to Ann and has two daughters: Jenny and Julia.

—Howard B. Lee

Further Reading

Bruno, J. E. (1996). Time perceptions and time allocation preferences among adolescent boys and girls. *Journal of Adolescence*, 31(121), 109–126.

James E. Bruno Research Activities page: <http://www.gseis.ucla.edu/faculty/bruno/brunor.htm>

Knowledge Factor Web site: <http://www.knowledgefactor.com/>

BUROS INSTITUTE OF MENTAL MEASUREMENTS

During the first quarter of the 20th century, testing became a big business in much of the developed world. Large numbers of authors and publishers started creating tests in fields such as education, psychology, and business to cater to an increasingly strong demand. Within about 25 years, more than 4,000 English-language tests were written and published. Many of these tests promised more than could be practically delivered and were formulated with only a vague understanding of the basic principles of measurement and statistics.

As a young professor of measurement and statistics at Rutgers University, Oscar K. Buros (1905–1978) became acutely aware of the insufficient technical

merits of many commercial tests. In the early 1930s, alarmed by the state of testing and seeking to improve the overall quality of tests, Buros began editing a series of books that featured descriptive test bibliographies and critical reviews written by recognized experts in the field. The results of these endeavors became the *Tests in Print* (TIP) and the *Mental Measurements Yearbook* (MMY) publication series. Buros used the term *mental measurements* to describe tests published in the broad areas of “aptitude, educational, intelligence, personality and psychological tests, questionnaires, and rating scales” (Buros, 1938, p. xiii). The Buros Institute of Mental Measurements was subsequently created as the organizational component to continue these publications.

Despite the pressing need for independent test evaluation, Buros initially found little financial support. By locating small grants, borrowing, and using his own funds, he was able to continue publication through the difficult years of the 1930s and 1940s. Financial pressures eased during the 1950s, when the TIP and MMY series became a recognized part of many university and public library collections.

After the death of Oscar Buros, his widow and long-time collaborator, Luella Gubrud Buros, moved the Institute to the University of Nebraska-Lincoln. Publication quickly resumed with *The Ninth Mental Measurements Yearbook* and *Tests in Print III*. At the present time, the Buros Institute is completing *The Seventeenth Mental Measurements Yearbook* (produced every 2 years) and *Tests in Print VIII* (produced every 3 years). In 2001, the Buros Institute introduced Test Reviews Online, a service providing ready access to a wide variety of information about and reviews of commercial testing instruments.

In order to fulfill the long-standing dream of Oscar Buros for an independent testing organization working to improve contemporary testing practices, the Buros Center for Testing was created in 1992. This new testing center brought together the products (MMY and TIP series) of the Buros Institute of Mental Measurements with the services (test oversight, evaluation, and independent psychometric research) of the newly created Buros Institute for Assessment Consultation and

Outreach. Together, these institutes continue the work of improving both the science and the practice of testing that were the primary focus of Oscar Buros's life.

—Robert A. Spies

Further Reading

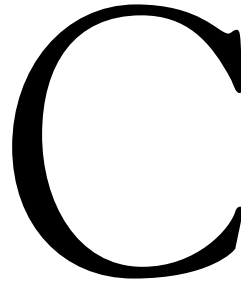
Buros, O. K. (Ed.). (1938). *The nineteen thirty eight mental measurements yearbook*. New Brunswick, NJ: Rutgers University Press.

Buros Institute Test Reviews Online: <http://www.unl.edu/buros/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from de la Rosa, I. A., Perry, J., Dalton, L. E., & Johnson, V. (2005). Strengthening families with first-born children: Exploratory story of the outcomes of a home visiting intervention. *Research on Social Work Practice*, 15(5), 323–338.

The **Buros Institute** is very well known for its collection of test reviews. Its online resources are the perfect place to begin looking for the appropriate measure. This study used the Buros Institute in its discussion of the Brigance Diagnostic Inventory of Early Development as a tool to evaluate the effectiveness of intervention, along with other measures. Using a theory-of-change framework, Iván A. de la Rosa and his colleagues examined outcome measures of a home visitation program that provided services to first-born children and their parents living in southwestern New Mexico. Home visitation workers conducted pretest and posttest assessments for prenatal and postpartum periods for 109 families receiving services in the First-Born Program. The results showed that clients participating in the First-Born Program displayed significantly higher posttest scores on measures of family resiliency. Specifically, clients demonstrated improved scores in operationalized measures of resilience: social support, caregiver characteristics, family interaction measures, and a reduction in personal problems affecting parenting.



An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.

—Max Planck

CALIFORNIA PSYCHOLOGICAL INVENTORY

The California Psychological Inventory (CPI; publisher: Consulting Psychologists Press) is a measure of normal personality and behavior. Originally developed in 1956 by Harrison Gough, the CPI provides a useful and accurate picture of people taking the instrument and a means for estimating their behavior across situations. The measure often is used in conjunction with assessing nonclinical populations and is appropriate for individuals age 13 years and older. The inventory takes approximately 45–60 minutes to complete.

Individual profiles are generated from the instrument based on its 20 folk concept scales. These scales are organized into four sectors: interpersonal style, intrapersonal style, achievement style and intellectual efficiency, and stylistic modes of thinking and behavior. The interpersonal style sector represents how individuals may be typed with regard to social interaction and includes the scales of dominance, capacity for status, sociability, social presence, self-acceptance, independence, and empathy. The intrapersonal style sector relates to an individual's values and self-regulation

and includes the responsibility, socialization, self-control, good impression, communality, well-being, and tolerance scales. The achievement style sector includes the achievement via conformance, achievement via independence, and intellectual efficiency scales. The stylistic modes sector includes psychological-mindedness, flexibility, and femininity/masculinity.

In addition to the 20 folk scales, the CPI includes 13 research and special purpose scales. These special purpose scales include managerial potential, work orientation, masculinity, femininity, anxiety, social desirability, acquiescence, creative temperament, leadership potential, amicability, law enforcement orientation, tough-mindedness, and narcissism. In addition to research, these scales are often used to explore occupational issues and are used frequently by organizations for such purposes as identifying leadership potential and managerial selection.

Underlying the scores on the folk concepts and special purpose scales are three structural scales that provide a measure of an individual's tendency toward or away from involvement with others, tendency to favor or doubt societal values, and perceived level of fulfillment or realization of abilities. The CPI also provides individual test takers with a description of how they would be described according to the

California Q-sort instrument. Further, the CPI provides a measure of the overall reliability and validity of each individual profile.

The CPI was last updated in 1996. The current (third) edition includes 434 items, 28 fewer than the preceding version. Items related to physical or psychological disabilities were removed for consistency with the 1990 Americans with Disabilities Act. Evidence of validity for the CPI is being collected continuously, and the instrument is used widely in counseling, organizational, and research settings.

—Todd J. Wilkinson and Jo-Ida C. Hansen

See also Minnesota Multiphasic Personality Inventory; NEO Personality Inventory; Sixteen Personality Factor Questionnaire

Further Reading

- Gough, H. G. (1996). *The California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- McAllister, L. W. (1996). *A practical guide to CPI interpretation* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Meyer, P., & Davis, S. (1992). *The CPI applications guide*. Palo Alto, CA: Consulting Psychologists Press.

Consulting Psychologists Press, Inc.: www.cpp-db.com

CAREER ASSESSMENT INVENTORY

The Career Assessment Inventory (CAI) is an interest inventory designed to survey an individual's interest in a variety of areas and then provide information on how those interests match up with the interests of people in a variety of occupations. Intended to assist individuals in career planning and decision making, it was authored by Charles B. Johansson and originally published in 1975. The CAI has undergone several revisions and is currently available in two versions: The Enhanced Version, for both college-bound and non-college-bound individuals, and the Vocational Version, for those who want to enter the workforce with little or no postsecondary training.

Both versions of the CAI use the widely accepted Holland model to organize information about general interest patterns into six general theme areas (realistic, investigative, artistic, social, conventional, and enterprising). Individuals are encouraged to develop a better understanding of their interest patterns so that they can use the information from the test when exploring occupations or other areas that are not specifically covered in the inventory. Two to eight Basic Interest Area Scales under each of the theme areas help individuals better understand their interest patterns; for instance, writing, creative arts, and performing and entertaining are under the artistic theme. While these patterns represent interests in specific areas, they are areas that could cover a wide variety of occupations. The greatest degree of specificity is found on the Specific Occupational Scales (111 for the Enhanced and 91 for the Vocational), which are each also organized into the general theme under which it best fits. These scales compare the responses of the individual taking the test with those of people in various occupations and indicate how similar they are.

Four scales are considered nonoccupational and measure one's orientation to learning by doing versus learning through traditional classroom work, introversion versus extroversion in the workplace, fine arts versus mechanical orientation, and the degree of variability of interests. These scales can be very useful in determining the validity of the test as well as in future planning.

The CAI takes 35–40 minutes to complete. Its 370 items are rated on a 5-point scale from *like* to *dislike*. It requires an eighth-grade reading level and is ideally suited for use with high school students who are in the process of career exploration, particularly if they are attempting to choose between attending college and entering the workforce directly. However, a number of other interest inventories are much more widely used (e.g., Strong Interest Inventory, Campbell Interest and Skill Survey, Self-Directed Search).

—Steve Saladin

See also Career Development Inventory; Career Maturity Inventory

Further Reading

Johansson, C. B. (1986). *Manual for the Career Assessment Inventory: The enhanced version*. Minneapolis, MN: National Computer Systems.

Career Assessment Inventory—The Enhanced Version:
http://www.pearsonassessments.com/tests/cai_e.htm

CAREER DEVELOPMENT INVENTORY

Career inventories can be separated into two categories, those that measure career choice content and those that measure career choice process. Inventories that deal with career choice content measure an individual's occupational abilities, vocational interests, and work values and then match these characteristics to the requirements, routines, and rewards that characterize a variety of occupations. Inventories that deal with career choice process measure an individual's attitudes toward, beliefs about, and competencies for making educational and vocational decisions as well as coping with vocational development tasks such as securing a job and establishing oneself in an organization. Process-oriented inventories provide a picture illustrating an individual's readiness and resources for career decision making. This picture portrays information regarding attitudes toward planning and exploration as well as knowledge about occupations and competence at decision making.

One popular instrument for measuring readiness and resources for educational and vocational decision making during adolescence is the Career Development Inventory (CDI). The CDI School Form is used with Grades 8 through 12, and the CDI College Form is used with college students.

The CDI is composed of two parts. Part I contains 80 items, takes approximately 40 minutes to complete, and reports scores for four scales: Career Planning (CP), Career Exploration (CE), Decision Making (DM), and Knowledge of the World of Work (WW). CP measures an individual's future orientation with regard to the world of work. Responses indicate the amount of thought an individual has given to future

occupational choices and the extent to which an individual has engaged in career planning activities. CE represents the degree to which an individual has made use of quality resources in career planning activities.

DM measures one's ability to apply the principles of rational decision making to educational and vocational choices. Brief scenarios describe individuals in the process of making career decisions. Based on the information given, the respondent must choose the most appropriate solution from a list of possible answers. It is proposed that individuals who can solve the career problems in these scenarios are likely to make wise decisions regarding their own careers. WW assesses one's knowledge regarding specific occupations and ways to attain, establish, and prosper in a job of one's own choosing.

Part II measures Knowledge of Preferred Occupation (PO). It contains 40 items and takes approximately 30 minutes to complete. Individuals are prompted to select their preferred occupational group from a list of 20 groups. The questions that follow address the type of work one should expect, educational requirements, values, and interests that are characteristic of individuals employed in that line of work.

The CDI reports three additional scores: Career Development Attitudes, which combines CP and CE; Career Decision Knowledge, which combines DM and WW; and Career Orientation Total, which combines CP, CE, DM, and WW. An online version of the CDI is available free to qualified professionals at <http://www.vocopher.com>.

—Kevin W. Glavin and Mark L. Savickas

See also Career Assessment Inventory; Career Maturity Inventory

Further Reading

Savickas, M. L. (1984). Career maturity: The construct and its measurement. *Vocational Guidance Quarterly*, 32, 222–231.

Savickas, M. L. (2000). Assessing career decision making. In E. Watkins & V. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed.; pp. 429–477). Hillsdale, NJ: Erlbaum.

CAREER MATURITY INVENTORY

Comprehensive models of vocational development during adolescence address the content of occupational choices and the process of career decision making. Choice content matches an individual's abilities and interests to occupational requirements and rewards. A good match leads to success and satisfaction, a poor match to failure and frustration. Career choice process deals with how individuals make decisions, not which occupation they choose. Individuals who apply a more highly developed and mature decisional process usually make more realistic and suitable choices. Because career maturity is central to adolescent vocational development, several inventories have been designed to measure *career choice readiness*. The Career Maturity Inventory, created by John O. Crites in 1978, was the first such measure to be published, and it subsequently became one of the most popular readiness inventories for students in Grades 6 through 12.

The Career Maturity Inventory measures the process dimension of vocational development during adolescence. The process dimension consists of two group factors: career choice attitudes and career choice competencies. Decision-making attitudes are viewed as dispositional response tendencies that mediate both choice behaviors and the use of the competencies. The decision-making competencies are viewed as comprehension and problem-solving abilities that pertain to making occupational choices. The attitudes are considered affective variables, and the competencies are considered cognitive variables.

The 1978 version of the Attitude Scale of the Career Maturity Inventory remains available in two forms: a screening form and a counseling form. The screening form consists of 50 items that yield a general score to indicate overall degree of career choice readiness. It is best used for performing needs analyses, evaluating career education interventions, and conducting research. The counseling form uses the same 50 items and adds 25 more. It yields the same general score as the screening form and also provides scores for five subscales: decisiveness, involvement, independence, orientation, and compromise. As its

name implies, it is best used during individual counseling sessions aimed at fostering vocational development and increasing career choice readiness. The 1978 version of the CMI also included five separate 25-item cognitive tests to measure the five decision-making competencies of self-appraisal, occupational information, goal selection, planning, and problem solving. Because of the 2.5 hours required to complete the five competency tests, few counselors or researchers ever used them.

Attempting to provide a briefer test, Crites and Mark L. Savickas constructed a 1995 version of the CMI that measures both the attitudes and the competencies. The CMI-R includes content appropriate for use with high school students as well as postsecondary adults. The CMI-R yields separate scores for its Attitude Scale, Competence Test, and Career Maturity Inventory. Five items from each of the 1978 counseling form subscales constitute the CMI-R Attitude Scale. The CMI-R Competence Test also consists of 25 items, five for each of the five competencies. The CMI-R total score merely sums the scores for the Attitude Scale and the Competency Test. The CMI screening form is available to qualified professionals free of charge at <http://www.vocopher.com>. The CMI-R is available from Western Educational Assessment, in Boulder, Colorado.

—Sarah A. Lopienski
and Mark L. Savickas

See also Career Assessment Inventory; Career Development Inventory

Further Reading

- Crites, J. O. (1978). *Theory and research handbook for the Career Maturity Inventory* (2nd ed.). Monterey, CA: CTB/McGraw-Hill.
- Crites, J. O., & Savickas, M. L. (1996). Revision of the Career Maturity Inventory. *Journal of Career Assessment*, 4, 131–138.
- Savickas, M. L. (1984). Career maturity: The construct and its measurement. *Vocational Guidance Quarterly*, 32, 222–231.
- Savickas, M. L. (2000). Assessing career decision making. In E. Watkins & V. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 429–477). Hillsdale, NJ: Erlbaum.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Hardin, E. E., & Leong, F. T. L. (2004). Decision-making theories and career assessment: A psychometric evaluation of the Decision Making Inventory. *Journal of Career Assessment*, 12(1), 51–64.

To address criticism that the empirical literature on assessment of career decision making lacks a theoretical base, the present study explored the relevance of a general theory of decision making to career decision making by assessing the psychometric properties of the Decision Making Inventory (DMI), designed to measure J. Johnson's decision-making styles, and by exploring the usefulness of the DMI as a predictor of career maturity. The DMI, the Attitudes Scale of the **Career Maturity Inventory** (CMI), and the Self-Concept Scale were administered to European American college students. The DMI demonstrated adequate reliability, the expected factor structure, and good convergent validity. Relationships with certain subscales of the CMI suggest the DMI has useful predictive validity. Similarities and differences between genders in the relationships between the DMI and CMI were found.

CARROLL DEPRESSION SCALE

The Carroll Depression Scale is one of three measures of depression developed by Bernard Carroll. The Carroll Depression Scales include the Carroll Depression Scale (CDS), the Carroll Depression Scale Revised, and the Brief Carroll Depression Scale. The CDS is published by Multi-Health Systems (www.mhs.com).

The CDS is a self-report inventory that takes 10–20 minutes to complete. It is used to measure depression symptomatology and symptom severity. The CDS is appropriate for adults age 18 years and older. There is no upper age limit for the CDS. The publishers suggest it is particularly useful for older adults and severely depressed people because of the cognitive simplicity of the yes-no response format.

The self-report items were designed to match the content of the Hamilton Depression Rating Scales. The CDS includes 52 items. The Revised CDS adds

9 items that assess subtypes of depression defined in the fourth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*. The Brief CDS consists of 12 of the 52 CDS items. Items assess central features of depression, such as appetite, energy, crying, and sexual interest.

Field trials for the CDS were conducted at the University of Michigan, Ann Arbor, in the 1970s; the CDS has also been used in trials at Duke University Medical Center. The participant pool from these two settings includes 959 depressed patients and 248 non-depressed people and has been used in a variety of reliability and validity studies. The participant pool is 100% White; other information about it is very limited. Given the test's norm group, it may have limited utility with non-White populations.

Validity data suggest good face, convergent, and discriminant validity. Face validity is demonstrated by a .80 correlation between scores from the CDS and scores from the Hamilton Depression Rating Scales. Convergent validity is demonstrated by moderate to high correlations with the Clinical Global Rating of Depression (.63), the Montgomery-Asberg Depression Rating Scale (.71), the Beck Depression Inventory (.86), and the Center for Epidemiological Studies of Depression Scale (.67). Discriminant validity is shown by the CDS's ability to differentiate depressed from anxious patients. The State-Trait Anxiety Inventory and the CDS correlate at a rate of .26.

Reliability data suggest strong split-half and test-retest reliability. Cronbach's alpha for the CDS is .95. Split-half reliability between odd and even items is .87. The Pearson correlation coefficient is .96. However, test-retest reliability was measured only on those patients whose Hamilton Depression Rating Scales scores did not vary more than two points between administrations. This restriction may have inflated reliability estimates. The authors state they chose to restrict the data in this way in their belief that the CDS is a state rather than a trait measure.

—Kathryn H. Ganske

See also Beck Depression Inventory; Clinical Assessment of Depression

Further Reading

- Carroll, B. J. (1981). The Carroll Rating Scale for Depression: I. Development, reliability, and validation. *British Journal of Psychiatry*, 138, 194–200.
- Feinberg, M., Carroll, B. J., & Smouse, P. E. (1981). The Carroll Rating Scale for Depression: III. Comparison with other rating instruments. *British Journal of Psychiatry*, 138, 205–209.
- Smouse, P. E. (1981). The Carroll Rating Scale for Depression: II. Factor analyses of the feature profiles. *British Journal of Psychiatry*, 138, 201–204.

Rating scales for mood disorders: <http://www.mhsource.com/disorders/diagdepress.html>

CATEGORICAL VARIABLES

A categorical variable is one that takes on values in a set of categories, as opposed to a continuous variable, which takes on a range of values along a continuum. The simplest examples of categorical variables are binary variables with only two possible responses, for instance “yes” and “no.” Categorical variables are most common in the social, biological, and behavioral sciences, although they can be found in almost any area of application. For example, the variable of marital status can be described as “single,” “married,” “divorced,” or “widowed”: four categories. The variable sex can be described as “male” or “female”: two categories. Education level can be classified as “grammar school only,” “some high school,” “completed high school,” “some university,” “completed university,” or “advanced or professional degree.”

When the categories ascribed to the variable are labels only, with no intrinsic ordering, then the variable is *nominal*. For example, it is generally meaningless to say that an individual who is divorced has higher or lower marital status than an individual who is widowed. Hence marital status is a nominal categorical variable. On the other hand, when the categories are naturally ordered, as with education level, socioeconomic status, or evaluation on a scale ranging from *strongly disagree* to *strongly agree*, then the variable is an *ordinal* categorical variable. In this case,

qualitative comparisons of individuals in different categories are meaningful. It is sensible to state that a person who has completed university has attained a higher level of education than a person who has completed only high school.

Categorical variables can be used as either the explanatory or the response variable in a statistical analysis. When the response is a categorical variable, appropriate analyses may include generalized linear models (for dichotomous or polytomous responses, with suitable link functions), log linear models, chi-square goodness-of-fit analysis, and the like, depending on the nature of the explanatory variables and the sampling mechanism. Categorical variables also fill a useful role as explanatory variables in standard regression, analysis of variance, and analysis of covariance models, as well as in generalized linear models. When the data are cross-classified according to several categorical variables (that is, when they come in the form of a table of counts), analyses for contingency tables, including log linear models, are appropriate. It is important to heed the distinction between nominal and ordinal variables in data analysis. When there are ordinal variables in a data set, the ordering needs to be entered explicitly into the model; this is usually achieved by incorporating constraints on parameters, which makes the analysis more complex.

In sum, categorical variables arise in a wide variety of scientific contexts. They require specialized statistical techniques, and these have been developed both theoretically and in terms of practical implementation.

—Nicole Lazar

See also Interval Level of Measurement; Nominal Level of Measurement; Ordinal Level of Measurement; Ratio Level of Measurement

Further Reading

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Fienberg, S. E. (1994). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.

CAUSAL ANALYSIS

Whether causal relationship can be established between two phenomena is highly controversial as a philosophical issue. According to David Hume, causal connections among real events cannot be perceived. Bertrand Russell even suggested that the term *causality* be banned in scientific discourse. Academic researchers, however, simply cannot help asking why and searching for causal explanations. Then Herbert Simon proposed that discussions of causality be restricted to our model of the reality rather than to the reality per se.

The desire to find causality is motivated by at least three potential benefits. First, causal explanation is believed to transcend time and space and therefore has a much wider scope of applications. Second, causal connections constitute the foundation of good control of the interested phenomena, which is especially important for policy making. Third, most causal statements are subject to the laws of logic and therefore more rigorous than they would be if no logic were involved.

It is unjustified, however, to think that causal analysis is always desirable and superior to other types of analysis. Sometimes, it is perhaps more important to know what has happened than why something has happened, because it is either less important or extremely difficult to ascertain the causal mechanism. Carefully produced and interpreted, descriptive statistics are sufficient in many situations. Furthermore, causality is not a prerequisite of precise prediction. For example, in demography and econometrics, making an accurate prediction is more important than identifying a causal chain or measuring a causal effect.

Finally, to *discover* that A causes B is beyond the capacity of statistical analysis. Causal mechanisms are either found through heuristic methods, such as experiment and observation, or simply derived from current theories and knowledge. It is only after the causal relationship is

proposed that statistical techniques can be applied to measure the size of the causal effect.

Basic Concepts

Causality implies a force of production. When we say A causes B, we mean the connection is a directed one, going from A to B. A causal relationship is thus asymmetric. Sometimes, the term *reciprocal causality* is used, meaning A causes B and B causes A as well. Although that relationship is possible, very often it indicates that a temporary order or an underlying causal mechanism has not been clearly identified.

In theory, causal relationships are best established when two conditions are satisfied. Call the causal variable *X* and the outcome variable *Y*, and suppose there are only two possible values for each of them: for the cause, “exists” or “not exists,” and for the outcome, “has happened” or “has not happened.” Then four scenarios are possible.

For example, consider whether taking a personal tutorial would improve a student’s exam results (Table 1). The tutorial does not help if the student takes it but no improvement follows (scenario 2) or if the student does not take it but improvement is made (scenario 3). We can say the tutorial is a causal factor of improvement only when scenarios 1 and 4 are *both* true, that is, if the student takes the tutorial and improvement follows *and* if not, no improvement. Researchers, however, are often tempted to make a causal connection when only scenario 1 is true, without seriously considering scenario 4. This is not only because it is straightforward to make the casual link from scenario 1 but also because information for scenario 4 is usually not available. A student either has taken the tutorial or has not taken it; it is impossible to have taken and have not

Table 1 Scenarios of Cause and Outcome

		Outcome (Y)	
		Has happened	Has not happened
Cause (X)	Exists	Scenario 1	Scenario 2
	Does not exist	Scenario 3	Scenario 4

taken a tutorial at the same time. Unless repeated measures are made at different times, either scenario 1 or scenario 4 is counterfactual. (Since the mid-1970s, statisticians and econometricians have developed a series of methods for estimating causal effects in counterfactual situations. The reader is referred to the second item in “Further Reading” below for details.)

Causal relationship in these analyses is probabilistic rather than deterministic. Deterministic causality applies to all cases under study without any exception. In the sciences of human beings, however, researchers can hope to measure only the probability of a causal connection for a particular case, or they have to work at an aggregate level. Exceptional deviant cases should not be taken as evidence of disproof.

The Commonly Used Three Criteria of Causality

Following John Stuart Mill, statisticians have identified three criteria for inferring a causal relationship: (a) covariation between the presumed cause and outcome; (b) temporal precedence of the cause; and (c) exclusion of alternative explanations for cause-outcome connections. All three have to be satisfied at the same time in order for causality to be derived.

The first is the easiest to establish—statistics such as the Pearson correlation coefficient, the score of chi-square, and the odds ratio are readily available for testing and measuring the covariation between two variables. As widely acknowledged, however, covariation alone cannot lead to causal statements.

The second criterion is more complicated and difficult to satisfy. The main problem lies in the uncertain relationship between temporal order and logical order. A cause must both temporally and logically precede an outcome, but a precedent event may or may not be a logical cause. Temporal connection is a special type of covariation and, therefore, does not necessarily lead to causal connection. For example, people acquire several fixed attributes at birth, such as sex, age, ethnicity, order among siblings, and so forth, but it makes little sense to take them as direct causes of other attributes that are developed at later stages of life, such as political affiliation.

The last criterion is said to be the most difficult—researchers can never be completely certain that all alternative explanations have been considered and excluded. The exclusion may be best handled in experiments because all the known possible causes have been kept “constant” so that their effects will not intervene in the causal relationship under study. But experiments are not immune to limitations: There is no way to ensure that *all* possible causes have been or can be taken into account; some variables, especially the fixed variables, or even a variable like education, cannot be manipulated; some participants comply with the experiment rules, but others do not; it may be unethical to conduct experiments on a certain group of participants; and finally, it may be difficult to prevent the participants from influencing one another.

In short, it is very difficult to establish causal relationships firmly, and the problem cannot be solved easily with experimental or longitudinal designs. This does not mean, however, that we should stop doing causal analysis. The logic holds for all statistical analyses—although we cannot prove that something is true, we can measure the *likelihood* that something is *not* true by analyzing the available information. If currently no information casts a serious doubt on the proposed causal relationship, we keep it. Further consistent results will increase the level of our confidence, while new inconsistent evidence will help us modify or even abandon the previous findings.

Causal Analysis in Nonexperimental Research: The Structural Equation Modeling Approach

Due to its aforementioned limitations, the experimental method is often infeasible for most social and behavioral studies. Causal analysis of nonexperimental (or observed) data (surveys, administrative records, etc.) does not aim to confirm causal relationships by manipulating the causal factors. Rather, the objective is to measure the causal effect and to disprove a hypothetical causal connection without any commitment to completely accepting a causal relationship.

The counterfactual approach, although statistically established, has not been widely applied in the social

sciences. A more widely followed approach is *structural equation modeling* (SEM). It has two main advantages. First, it combines *path analysis* and *factor analysis*. Path analysis enables us to statistically test causal relations among a set of observed variables, but it does not deal with variables that are not directly measurable (latent variables), such as socioeconomic status, social capital, intelligence, and so on. Factor analysis enables us to link some observed variables to a latent variable, but it is not designed to test causal relations among these variables. SEM is a powerful method that measures causal relations among both observed and latent variables. Second, measurement errors are assumed marginal and thus ignorable in many statistical techniques. In contrast, SEM explicitly incorporates the error terms into a statistical model, generating a more reliable measurement of causal relationship with the measurement errors corrected. The reader who would like to learn more can start with the first reference under “Further Reading.”

—Keming Yang

See also Structural Equation Modeling

Further Reading

- Saris, W. E., & Stronkhorst, H. (1984). *Causal modelling in nonexperimental research: An introduction to the LISREL approach*. Amsterdam: Sociometric Research Foundation.
- Winship, C., & Sobel, M. (2004). Causal inference in sociological studies. In M. Hardy and A. Bryman (Eds.), *Handbook of data analysis* (pp. 481–503). London: Sage.

LISREL student version (8.7) for downloading: <http://www.ssicentral.com/index.html> (Note that there are limitations on the size of the model and no technical support.)

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Henry, S. (2000). What is school violence? An integrated definition. *Annals of the American Academy of Political and Social Science*, 567(1), 16–29.

In this study, **causal analysis** is used to address the wider context of school violence, the wider forms of violence in schools, and the important interactive and causal effects arising from the

confluence of these forces. Usually, such studies focus on interpersonal violence between students or by students toward their teachers. Stuart Henry argues that not only does the complexity of this issue defy such a simplistic framing but dealing with the problem at that level does not go far enough. What is needed is an integrated, multi-level definition of the problem that will lead to a multilevel causal analysis and a comprehensive policy response that takes account of the full range of constitutive elements. Here, the first stage of such an approach is outlined with regard to defining the nature and scope of the problem.

CENSORED DATA

Suppose that in an experiment or study, a group of individuals (objects, patients, or devices) is followed over time with the goal of observing an event such as failure or death. Individuals who do not experience the event of interest in the observation period are said to be censored, and the data obtained from such individuals are known as censored data.

In most cases, experiments or studies have a finite observation period, so for some individuals, the observation period may not be long enough to observe the event of interest. Also, individuals may cease to be at risk before the observation period. For example, in a clinical trial setting, patients might drop out of the study, or in the case of testing the reliability of a device, the device may fail for reasons other than the one the experimenter is interested in. Such individuals known not to experience the event within or before the observation period are said to be right censored.

Censored data may demonstrate left censoring and interval censoring, as well. In the former, the event of interest occurs before the observation period. For example, suppose a group of women is selected to be followed for the development of breast cancer. If some of these women had already developed breast cancer, then the time to the development of breast cancer is left censored for them. In the case of interval censoring, one observes an interval within which the event of interest occurred, but the actual time of

occurrence remains unknown. Interval censoring occurs when devices are tested only at specific times, say $t_1, t_2, \dots, t_k$, and failures occur between two consecutive times. Right and left censoring are special cases of interval censoring with the intervals (T, ∞) and $(0, S)$, respectively, where S is the starting time and T is the ending time of the study.

When data contain censored observations, special care has to be taken in the analysis. Common statistical methods used to analyze censored data include the Kaplan-Meier estimator, the log-rank test, the Cox proportional hazard model, and the accelerated failure time model.

—Abdus S. Wahed

See also Observational Studies

Further Reading

- Kalbfleisch, J. D., & Prentice, R. L. (2002). *Statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.
- Walpole, R. E., Myers, R. H., & Myers, S. L. (1998). *Probability and statistics for engineers and scientists* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

CENTERS FOR DISEASE CONTROL AND PREVENTION

The Centers for Disease Control and Prevention (CDC), a unit of the Department of Health and Human Services, is a U.S. government public health agency with a workforce of almost 6,000 persons currently under the direction of Dr. Julie Louise Gerberding. Headquartered in Atlanta, Georgia, CDC has 10 other locations in the United States and Puerto Rico. CDC's mission encompasses several goals: (a) to protect the public's health and safety; (b) to educate the public through dissemination of reliable scientific health information; (c) to prevent and control disease, injury, and disability; and (d) to establish strong partnerships with numerous public and private entities such as local and state health departments, academic institutions, and international health organizations.

Following World War II, Dr. Joseph W. Mountin formed the Communicable Disease Center on July 1, 1946, in Atlanta as a peacetime infectious disease prevention agency based on the work of an earlier agency, the Malaria Control in War Areas. CDC's original focus was the problems of malaria and typhus, later broadening to diseases such as polio (1951) and smallpox (1966). In 1970, the agency became known as the Center for Disease Control to reflect its broader mission. In 1992, it added the term *prevention* to its name but remained known as CDC.

CDC is now organized into six coordinating offices for Global Health, Terrorism Preparedness and Emergency Response, Environmental Health and Injury Prevention, Health Information and Services, Health Promotion, and Infectious Diseases. These coordinating offices are further divided into 12 centers, each of which has its own areas of expertise and public health concerns. For instance, the National Center for Injury Prevention and Control aims to reduce mortality, disability, and costs related to injuries resulting from events such as motor vehicle accidents, youth violence, child maltreatment, and suicide. The National Center for Health Statistics documents the health status of the United States population, monitors trends in health care delivery and utilization, and evaluates the impact of health policies and programs through statistical computations. The Office of the Director is responsible for the management, oversight, and coordination of the scientific endeavors of all centers.

CDC's working budget for the fiscal year 2005 was estimated at \$7.7 billion, with the highest appropriation assigned to efforts to combat HIV and AIDS. From its beginnings in 1946, with a budget of less than \$10 million, CDC has become the nation's premiere public health agency with the stated vision for the 21st century of healthy people in a healthy world through prevention.

—Marjan Ghahramanlou-Holloway

Further Reading

- Etheridge, E. W. (1992). *Sentinel for health: A history of the Centers for Disease Control*. Berkeley: University of California Press.

Snider, D. E., & Satcher, D. (1997). Behavioral and social sciences at the Centers for Disease Control and Prevention: Critical disciplines for public health. *American Psychologist*, 52(2), 140–142.

Department of Health and Human Services—Centers for Disease Control and Prevention: www.cdc.gov

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Whitaker, D. J., Lutzker, J. R., & Shelley, G. A. (2005). Child maltreatment prevention priorities at the Centers for Disease Control and Prevention. *Child Maltreatment*, 10(3), 245–259.

The **Centers for Disease Control and Prevention** (CDC) is the United States' public health agency and deals with many different types of public health issues. The Division of Violence Prevention at CDC's National Center for Injury Prevention and Control has had a long-standing interest in the prevention of child maltreatment. As the nation's public health agency, CDC seeks to focus the public health perspective on the problem of child maltreatment and to promote science-based practice in the field. Since 1999, CDC has developed research priorities to address the prevention of child maltreatment. This article provides a brief rationale for applying a public health approach to child maltreatment and a discussion of the priority-setting process, priorities in each of four areas of the public health model, and some of CDC's current child maltreatment prevention activities.

relatively independently occurring variables, the attribute will be normally distributed. Such attributes are ordinarily complex variables that require considerable time in their development. For instance, a multitude of relatively independently occurring variables influence adult reading achievement, each of which may be influential in large or small ways:

1. Quality of teachers in the early grades
2. Variety of reading matter available (books, magazines, newspapers)
3. Parental encouragement to read
4. Genetic endowment
5. Diet during developmental years
6. Socioeconomic level of the child's family
7. Number of competing activities available (e.g., arcade games, sports, television)
8. Interests and activities of friends
9. Initial success (or failure) in reading activities
10. Physical athleticism of the child

It is easy to imagine many other such variables. To the extent that the variables produce small, relatively random effects on adult reading achievement over time, the distribution of adult reading achievement will be normally distributed. Because the variables that influence such attributes produce unrelated effects, one's position in the distribution of scores in the population is largely a matter of luck. Some individuals are very lucky (e.g., they may have parents who encourage reading, access to reading materials, excellent early teachers) and enjoy high adult reading achievement. Some are very unlucky (e.g., they may have impoverished parents who do not support reading, poor diet, frequent illnesses) and remain illiterate. But about two thirds are neither very lucky nor very unlucky and become average adult readers. The fuzzy central limit theorem then, provides an explanation for the relatively common occurrence of normal distributions in nature.

The second reason the central limit theorem is important concerns its use in statistical inference. Because in a normal distribution we know the

CENTRAL LIMIT THEOREM

The central limit theorem states that, under conditions of repeated sampling from a population, the sample means of random measurements tend to possess an approximately normal distribution. This is true for population distributions that are normal and decidedly not normal.

The central limit theorem is important for two reasons. The first can be understood by examining the *fuzzy central limit theorem*, which avers that whenever an attribute is influenced by a large number of

percentage of cases falling between any two points along the baseline *and* the percentage of cases above and below any single point along the baseline, many useful inferences can be made. Consider an investigation comparing two treatment approaches designed to enhance math achievement. Following the treatment, the dependent variable, a test of math achievement, is administered. By comparing the two treatment group means, the researchers determine the probability that the difference between the means could have occurred by chance if the two treatment groups were drawn from the same population distribution. If the probability were, say, 0.0094, the investigators could have a high degree of confidence in rejecting the null hypothesis of no difference between the treatments. If the central limit theorem did not commonly apply in nature, such inferences could not be made.

—Ronald C. Eaves

See also Sampling Distribution of a Statistic

Further Reading

Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Author.

Sampling distributions and the central limit theorem:
http://people.hofstra.edu/faculty/Stefan_Waner/RealWorld/finitetopic1/sampldistr.html

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Essex, C., & Smythe, W. E. (1999). Between numbers and notions: A critique of psychological measurement theory & psychology. *Theory & Psychology*, 9(6), 739–767.

The **central limit theorem** plays a critical role in statistics. This article discusses the applications of mathematical machinery to psychological ideas and how that machinery imposes certain requirements on the relationship between numbers and notions. These imposed choices are driven by the mathematics and not the psychology. Attempting a theory-neutral approach to research in psychology, where commitments in response to the options are made unknowingly, becomes instead

a theory-by-default psychology. The article begins to catalog some of these mathematical choices to make them explicit in order to allow psychologists the opportunity to make explicit theoretical commitments.

CENTROID

The notion of centroid generalizes the notion of a mean to multivariate analysis and multidimensional spaces. It applies to vectors instead of scalars, and it is computed by associating to each vector a mass that is a positive number taking values between 0 and 1 such that the sum of all the masses is equal to 1. The centroid of a set of vectors is also called the *center of gravity*, the *center of mass*, or the *barycenter* of this set.

Notations and Definition

Let $\mathcal{V}$ be a set of I vectors, with each vector being composed of J elements:

$$\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_i, \dots, \mathbf{v}_I\} \text{ with } \mathbf{v}_i = [v_{i,1}, \dots, v_{i,j}, \dots, v_{i,J}]^T.$$

To each vector is associated a mass denoted m_i for vector i . These masses take values between 0 and 1, and the sum of these masses is equal to 1. The set of masses is a vector denoted $\mathbf{m}$. The centroid of the set of vectors is denoted $\mathbf{c}$, defined as

$$\mathbf{c} = \sum_i^I m_i \mathbf{v}_i.$$

Examples

The mean of a set of numbers is the centroid of this set of observations. Here, the mass of each number is equal to the inverse of the number of observations: $m_i = \frac{1}{I}$.

For multivariate data, the notion of centroid generalizes the mean. For example, with the following three vectors,

$$\mathbf{v}_1 = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 8 \\ 24 \end{bmatrix}, \text{ and } \mathbf{v}_3 = \begin{bmatrix} 16 \\ 32 \end{bmatrix},$$

and the following set of masses,

$$m_1 = \frac{1}{2}, \quad m_2 = \frac{1}{8}, \quad \text{and} \quad m_3 = \frac{3}{8},$$

we obtain the following centroid:

$$\begin{aligned} \mathbf{c} &= \sum_i m_i \mathbf{v}_i = \frac{1}{2} \begin{bmatrix} 10 \\ 20 \end{bmatrix} + \frac{1}{8} \begin{bmatrix} 8 \\ 24 \end{bmatrix} + \frac{3}{8} \begin{bmatrix} 16 \\ 32 \end{bmatrix} \\ &= \begin{bmatrix} 12 \\ 25 \end{bmatrix}. \end{aligned}$$

In this example, if we plot the vectors in a two-dimensional space, the centroid would be the center of gravity of the triangle made by these three vectors with masses assigned proportionally to their vector of mass. The notion of centroid can be used with spaces of any dimensionality.

Properties of the Centroid

The properties of the centroid of a set of vectors closely parallel the more familiar properties of the mean of a set of numbers. Recall that a set of vectors defines a multidimensional space, and that to each multidimensional space is assigned a generalized Euclidean distance. The core property of the centroid is that the centroid of a set of vectors minimizes the weighted sum of the generalized squared Euclidean distances from the vectors to any point in the space. This quantity that generalizes the notion of variance is called the *inertia* of the set of vectors relative to their centroid.

Of additional interest for multivariate analysis, the *theorem of Huyghens* indicates that the weighted sum of the squared distances from a set of vectors to any vector in the space can be decomposed as a weighted sum of the squared distances from the vectors to their centroid plus the (weighted) squared distance from the centroid to this point. In term of inertia, Huyghens's theorem states that the inertia of a set of vectors to any point is equal to the inertia of the set of vectors to their

centroid plus the inertia of their centroid to this point. As an obvious consequence of this theorem, the inertia of a set of vectors to their centroid is minimal.

—Hervé Abdi

See also Correspondence Analysis; Discriminant Correspondence Analysis; Distance; DISTATIS; Multiple Correspondence Analysis; Multiple Factor Analysis; STATIS

CHANCE

The word *chance* originated in the Latin word for “to fall.” Chance occurs the way things just happen to fall. Like a coin toss, chance events occur unpredictably, without any apparent or knowable causes, or by accident—the latter word deriving from the same Latin root. Although *chance* is a very useful word in everyday language, its status as a term in measurement and statistics is much more ambivalent. On the one hand, the word is almost never treated as a technical term. Thus, it is seldom given a precise definition, or even an intuitive one. Indeed, it is rare to see *chance* as an entry in an index to any book on measurement and statistics.

On the other hand, both the concept and the word permeate articles and books on these subjects. It is especially hard to imagine writing a statistics textbook without having recourse to the word on numerous occasions. Furthermore, the word *chance* is used in a wide variety of ways. In particular, we may distinguish chance as probability, as unknown determinants, as technique, and as artifact.

Chance as Probability

What is now called *probability theory* had its origins in a famous exchange of letters between the mathematicians Pierre de Fermat and Blaise Pascal concerning games of chance. As a consequence, early mathematical treatments would be identified as the “logic of chance” (John Venn) or the “doctrine of chance” (Abraham de Moivre). But by the time Pierre-Simon Laplace published his classic *Analytical*

Theory of Probabilities, the word *chance* had seemingly ceased to have scientific content, thereby returning to the status of a mundane word. It had been replaced by the word *probability*.

Despite this shift in significance, *chance* is still used to refer to the probability of an event or set of events—whether or not they concern outcomes in games of chance. When used in this manner, *chance* is often treated as something tantamount to a rough measure of how much we should anticipate an event's occurring. Some events may have a good chance of happening, others a very poor chance, yet others a middling chance. Two events can also be said to have an equal chance of happening (e.g., the balls in urn models all have an equal chance of being chosen). Here, chance functions as a generic term that encompasses more specialized and precise concepts, including an event's probability (the number of times an event happens divided by the total number of times it could have happened) and an event's odds (the probability of an event occurring divided by the probability of an event not occurring).

One could argue that it is no longer proper to use such a vague word when more-precise terms are readily available. Yet the very imprecision of *chance* can prove to be an asset in certain contexts. In particular, *chance* becomes serviceable when a more precise term is unnecessary to convey a statistical idea. For example, a simple random sample can be defined as a sample in which each case in the entire population has an equal chance of being selected for inclusion. Substituting "probability" for "chance" may not necessarily improve this definition, especially given that the actual probability may not be known or even calculable.

Chance as Unmeasured Determinants

Chance is often assumed to operate as an active agent that helps account for a particular pattern of events. More specifically, chance is frequently invoked as an unknown force or set of forces that explains why certain expectations may be violated. A simple example can be given from classical test theory. A research participant's response on a particular test item may not

accurately reflect the participant's actual ability, attitude, or disposition. Instead of the *true score*, we obtain a *fallible score*. According to classical theory, the fallible score is the sum of the true score plus error. The error may encompass a large number of extraneous factors, such as momentary inattentiveness, a careless recording mistake, a misunderstanding of the question, a classically conditioned emotional response, and so forth. Significantly, these causes are assumed to be so numerous and mutually independent that the error can be considered pure chance. Hence, the fallible score is the sum of the true score and unmeasured chance determinants. This assumption is absolutely crucial in the classical theory of measurement. It means that the error is uncorrelated with the true score for that item or any other item, and that the error is equally uncorrelated with the errors of other items. As a consequence, a composite score defined as the sum of fallible scores on a large number of items will contain much less error than the individual items because the separate errors will cancel each other out rather than accumulate. In other words, if errors are determined by chance, then a multi-item test will be much more reliable than any of the items that make it up.

Various statistical techniques also introduce a chance component, but with a different purpose in mind. For example, the dependent variable (or criterion) in a multiple regression equation can be expressed as a function of an intercept (or constant term), the weighted sum of the independent variables (or predictors), and an error term that represents the discrepancies between predicted and observed scores (i.e., the residuals). When regression is applied to correlational data, unbiased least-squares estimators of the weights for the independent variables (i.e., the unstandardized regression coefficients) can be obtained only if this error (or disturbance) term essentially constitutes a "random shock" consisting of pure chance. Here chance consists of a conglomerate of all those unspecified determinants of the dependent variable that are uncorrelated with independent variables specified in the regression equation.

This chance disturbance plays an even bigger role in time series analysis. In particular, this term is

essential to the very definitions of autoregressive and moving-average processes. In a first-order autoregressive process, for example, the value of a variable at time t is specified as a function of the value of the variable at time $t-1$ plus a random shock. Even more striking, in a first-order moving-average process, the value of a variable at time t is specified as a function of the value of a random shock at t minus a portion of the random shock at $t-1$ (i.e., the variable consists entirely of concurrent and lagged chance inputs).

Chance as Technique

Many common methodological procedures use chance to obtain certain desirable outcomes. An example was given earlier, namely, the use of random sampling to ensure that the sample is representative of the larger population. Moreover, knowing that each case has an equal chance of entering the sample permits the calculation of statistics that would not be available otherwise, such as the standard error of the mean. Other sampling strategies, such as stratified random sampling and probability sampling, also use chance to attain specific methodological ends. In probability sampling, for instance, each case in the population has a “known chance” (rather than an equal chance) of being selected for inclusion in the sample, a stipulation that still permits the inference of population attributes from sample characteristics.

Another technical use of *chance* is illustrated by experimental research in which the participants are randomly assigned to the various conditions. Such randomization plays an extremely important role in causal inference and has direct consequences for the statistical analyses. Previously it was noted that multiple regression assumes that all unmeasured causes of variation in the dependent variable are contained in an error term. Because the latter term must be uncorrelated with the independent variables in the equation, the error is assumed to be random. Whereas in correlational studies this randomness can only be assumed, in experimental studies the random assignment itself guarantees that the error or residual term represents pure chance. As a result, the estimated effects of the experimental manipulations are far more likely to be

unbiased than would be the situation if the conditions were chosen by the participants themselves.

A final example of chance as technique may be found in Monte Carlo methods. Named after the famed casino of Monaco, this approach is specifically inspired by games of chance. A roulette wheel will generate a series of random numbers, and Monte Carlo methods rely on random numbers to simulate processes that are assumed to possess a substantial chance component. Widely used in the natural sciences, Monte Carlo methods also have a significant place in measurement and statistics. When the properties of measures, techniques, or estimators cannot be determined analytically, suitable data can be simulated to test the properties empirically. For example, in studies of exploratory factor analysis, Monte Carlo methods have frequently been used to address such issues as the best factor extraction algorithms, the optimal decision criteria for determining the number of factors, and the most effective rotational schemes. Often these simulations take advantage of programs that generate random numbers with a specified distribution, most commonly a normal distribution (i.e., normal random deviates).

Chance as Artifact

With rare exceptions, such as researchers working with census data, empirical inquiries are most frequently carried out using relatively small samples from a larger population of cases. The characteristics of those samples will necessarily depart from the characteristics of the greater population. Hence, even if the means for two groups (e.g., men and women) are absolutely identical in the population, the means will differ in the sample. Similarly, even if the correlation between two variables was exactly zero in the population, the correlation may be substantially greater or less than zero in the sample. The smaller the sample size, or N , the more extensive the probable error. In fact, when N is very small, say less than a dozen cases, seemingly substantial mean differences or correlations can appear—statistical outcomes that still should be attributed to chance rather than to bona fide effects.

The most common solution to this problem is to implement null hypothesis significance testing. Typically, the null hypothesis is that mean difference, correlation, or another statistic is absolutely zero in the greater population. Using some statistical test, the researcher determines whether the sample values exceed what could reasonably be expected from chance fluctuations alone. So when we conclude that a result is significant at the .05 probability level, we are asserting that the chances are only 5 out of 100 that we could obtain an effect of that magnitude by sampling error. By rejecting the null hypothesis, we are saying that the observed mean difference or correlation is likely not a mere fluke of our little sample and that there is a very high probability that these statistics are nonzero in the larger population.

Unfortunately, this procedure cannot completely solve the problem. Sometimes researchers will adopt measurement strategies or statistical techniques that cause them to “exploit chance,” to take advantage of sampling error rather than reduce its impact. For instance, when item analysis is used to select the best items for inclusion in a multi-item test, the investigator naturally picks those items that correlate most highly with the overall score. Yet unless the sample size is large, the relative sizes of the item-composite correlations will be contaminated by considerable sampling error. As a result, the final test may be more representative of the sample than of the population. Some items will be incorrectly included while others are just as incorrectly excluded.

A similar problem arises in a particular form of multiple regression, namely, analyses employing a stepwise procedure for variable inclusion. In forward stepwise regression, for example, predictors are added to the equation one by one. At each step, that variable is inserted which makes the greatest improvement in the equation’s predictive power. Potential predictors that make no contribution to the explained variance are then omitted. Regression equations constructed in this way are taking inappropriate advantage of chance fluctuations in sample characteristics. Accordingly, the predictors included in the final equation may not completely replicate in other samples drawn from the same population. Indeed, in small samples, the equation

can contain predictors that are almost arbitrary, and the chances of complete replication become virtually zero.

It should be made explicit that the emergence of these chance artifacts is not simply a matter of sample size. Opportunities for exploiting chance also increase as we increase the number of parameters to be estimated (e.g., mean differences, correlations, regression coefficients). For instance, the more correlations computed from the data, the higher is the probability that a coefficient will emerge that is significant at the .05 level, even when all coefficients are zero in the population. Thus, increasing the number of correlations to be estimated increases the chance of rejecting the null hypothesis when the null hypothesis is in fact true (i.e., Type I error). Avoiding this unfortunate repercussion requires that the investigator implement procedures to correct the probability levels (e.g., the Bonferroni correction).

Another example also comes from multiple regression analysis. The more independent variables there are in the equation, the more parameters that must be estimated (viz. regression coefficients). Therefore, for a given sample size, equations with many predictors will have more opportunities to exploit chance than will equations with few predictors. This difference will manifest itself in the “squared multiple correlation” (R^2 , where R is the correlation between predicted and observed scores). This statistic will be biased upward by the opportunistic assignment of large regression weights to those independent variables that have their effects inflated by sampling error. To handle this adverse consequence, most statistical software publishes an “adjusted R^2 ” along with the regular squared multiple correlation. The adjustment makes allowance for the number of regression coefficients estimated relative to the sample size.

Clearly “chance as artifact” has a very different status from chance as probability, as unmeasured determinants, and as technique. The latter three usages take chance as a “good thing.” *Chance* provides a generic term for probability, odds, or likelihood, makes measurement and prediction errors less problematic, and offers useful tools for designing research and evaluating statistical methods. Yet the

intrusion of sampling error, along with the implementation of methods that accentuate rather than reduce that error, shows that chance can also be a “bad thing” that complicates analyses. Nonetheless, it should be evident that *chance* remains a very useful word despite its varied applications and vague definitions. It provides a catch-all term that can encompass a diversity of issues that are crucial to measurement and statistics.

—Dean Keith Simonton

See also Bonferroni Correction; Classical Test Theory; Exploratory Factor Analysis; Least Squares, Method of; Markov Chain Monte Carlo Methods; Null Hypothesis Significance Testing; Pascal, Blaise; Probability Sampling; Random Numbers; Random Sampling; Regression Analysis; Residuals; Sampling Error; Significance Level; Statistical Significance; Type I Error

Further Reading

Everitt, B. (1999). *Chance rules: An informal guide to probability, risk, and statistics*. New York: Springer.

Chance magazine: <http://www.amstat.org/publications/chance/>
Chance Web site devoted to the teaching of courses dealing with chance: <http://www.dartmouth.edu/~chance/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Oldman, D. (1974). Chance and skill: A study of roulette. *Sociology*, 8(3), 407–426.

If scientists have any job, it is trying to reduce the role that chance, or random variation, plays in scientific research. **Chance** is a ubiquitous factor in all our daily activities, and trying to explain how chance is addressed and dealt with is a central point in the study of measurement and statistics. This article takes an applied approach to the topic and focuses on how people handle random events. Many accounts of the chance element in games and many attempts at general explanations of gambling assume that an individual accepts the events as unpredictable and passively awaits the outcome, but a study of people playing roulette in a gaming club suggests that this is far from the case. Playing roulette can be seen as an exercise in “skill” that depends on the construction and

maintenance of predictive theories. One form of theorizing attributes causal efficacy to the croupier, and the game becomes a contest between croupier and player. This behavior is reinforced as players attempt to manipulate their working conditions and status. Players may then adopt a nonarithmetic calculus of win and loss that confirms their theorizing.

CHI-SQUARE TEST FOR GOODNESS OF FIT

The chi-square (pronounced “kai square” and often written as χ^2) family of statistical tests comprises inferential tests that deal with categorical data. In many situations, the researcher may not have data in the form of scores or measurements (ordinal, interval, or ratio level of measurement) for the cases in their sample but instead will have information concerning some classification (categorical level of measurement). For instance, if we were interested in smoking behavior, we may want to know how many cigarettes the average person smokes per day (interval data); alternatively, we may want to look more simply at whether people smoke or not (categorical data).

Typically, when we have scores or interval measurements, we are interested in the mean or the median score (the mean number of cigarettes smoked). With categorical data, however, it would be fairly pointless to calculate the mean (the average of the sample is a semismoker?). Instead, we are interested in the frequency of cases that come up in each classification of the variable we are interested in (how many people smoke and how many do not?). Chi-square tests can look at situations like this, in which cases are categorized on one factor (the chi-square goodness of fit), and also at cases categorized on two factors (sometimes referred to as being *cross categorized*), such as a smoker and an undergraduate (the chi-square measure of association). This entry focuses on the goodness-of-fit test.

Karl Pearson developed the goodness-of-fit test to determine whether observed distributions of frequency

data matched, or “fitted,” a theoretical distribution. By estimating the cell frequencies one would expect to observe under some theoretical distribution, one could compare the actual frequencies and calculate the difference. The smaller the difference, the better the fit.

The most common use of the chi-square is to compare the observed distribution with the distribution expected under the null hypothesis. The null hypothesis states that categorization will be random, and thus the expected cell frequencies will be based on chance. We can calculate the expected cell frequencies by dividing the number of cases in the sample by the number of possible classifications. If these expected frequencies are very similar to the observed frequencies, then the fit between the observed distribution and the null hypothesis distribution will be good. If they are quite different, then the fit will not be good, and we may reject the null hypothesis.

Using the test statistic and the degrees of freedom, we can estimate the significance value. Degrees of freedom for a goodness-of-fit test are calculated as the number of classifications minus 1.

The following assumptions underlie the test:

1. The data must be treated as categorical.
2. The categories must be mutually exclusive. This means that it must be impossible for a case to be in more than one category. For instance, a person is either a smoker or not.
3. None of the expected values may be less than 1.
4. No more than 20% of the expected values may be less than 5.

An Example

Suppose we asked 100 people to name their favorite season. The null hypothesis would state that there would be a chance distribution of responses, meaning that there would be a fairly even number of people selecting each season. So the expected frequencies under the null hypothesis would be calculated as the number of cases (100) divided by the number of possible classifications (4). Therefore, we would expect, under the null hypothesis, to observe 25 people

choosing spring as their favorite season, 25 choosing summer, 25 choosing autumn, and 25 choosing winter. Now we compare this to the frequencies we actually observe, shown in Table 1.

In Table 1, we see that the distribution does not look like a chance distribution. More people chose summer as their favorite season than chose any other season. To see if the null hypothesis model fits the observed data, we look at the difference between the expected cell frequencies and the observed cell frequencies (these differences are called the *residuals*). The sum of these residuals represents the goodness of fit. Before they can be summed, however, the residuals must be squared to avoid positive residuals cancelling out negative residuals. Chi-square is calculated using this equation:

$$\chi^2 = \Sigma([o - e]^2/e)$$

where e is an expected cell frequency and o is an observed cell frequency.

So for our example,

$$\chi^2 = (-4^2/25) + (23^2/25) + (-9^2/25) + (-10^2/25)$$

$$\chi^2 = 29.04.$$

Now we have to compare this value to the distribution of chi-square to assess significance. In this case, there are 3 degrees of freedom (because we have four observed residuals and have to subtract one degree of freedom for the model). Looking at a table of critical values for chi-square, we see that if the observed value is greater than 11.34, then it is significant ($p < .01$); because our observed value of chi-square (29.04) is greater than 11.34, we reject the null hypothesis.

Table 1 Hypothetical Data to Show the Frequency of People Choosing Each Season as Their Favorite

	Spring	Summer	Autumn	Winter
Expected	25	25	25	25
Observed	21	48	16	15
Residual	-4	23	-9	-10

Calculating Chi-Square Goodness of Fit Using SPSS

The chi-square statistic may be calculated in several ways using SPSS. The procedure described here is fairly straightforward for data inputted as per the screenshot in Figure 1.

1. Tell SPSS that the “count” variable is a frequency and not a score: Go to Data → weight cases and click next to “Weight cases by” and then put “count” into the “Frequency variable” box.
2. Go to Analyze → Nonparametric tests → Chi-square . . .
3. Put “count” into the “Test variable list” and click on “OK.”

The output will be a table of the descriptive statistics as well as a table that contains the chi-square value, the degrees of freedom, and the p value.

—Siân E. Williams

See also Chi-Square Test for Independence

Further Reading

Chi-square calculator: <http://faculty.vassar.edu/lowry/csfit.html>

	Season	Count	var	var	var	var	var	var	var	var	var	var	var
1	Spring	21.00											
2	Summer	48.00											
3	Autumn	16.00											
4	Winter	15.00											
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													
24													
25													
26													
27													
28													
29													
30													

Figure 1 Screenshot of Data Inputted for a Chi-Square Goodness-of-Fit Test

CHI-SQUARE TEST FOR INDEPENDENCE

The chi-square test for independence is a significance test of the relationship between categorical variables. This test is sometimes known as the “Pearson’s chi-square” in honor of its developer, Karl Pearson. As an example of this test, consider an experiment by David M. Lane, S. Camille Peres, Aniko Sándor, and H. Al Napier that evaluated the effectiveness of a new method for initiating computer commands. One group of participants was tested using a mouse; a second group was tested using a track pad. Although for both groups, the new method led to faster performance than did the standard method, the question addressed here is whether there was any relationship between preference for the new method and the type of pointing device (mouse or track pad).

Table 1 is a *contingency table* showing whether preference is contingent on the device used. As can be seen in Table 1, 4 of the 12 participants in the mouse group (33%), compared with 9 of the 10 participants (90%) in the track pad group, preferred the new method. Therefore, in this sample, there is an association between the pointing device used and the method preferred. A key question is whether this association in the sample justifies the conclusion that there is an association in the population.

The chi-square test for independence, as applied to this experiment, tests the null hypothesis that the preferred method (standard or new) is independent of the pointing device used (mouse or track pad). Another way of stating this null hypothesis is that there is no association between the categorical variables of preferred method and pointing device. If the null hypothesis is rejected, then one can conclude that there is an association in the population.

Calculations

The first step in the calculation is to find the expected frequencies in each cell

Table 1 Data From the Example Experiment

<i>Device</i>	<i>Preferred Method</i>		<i>Total</i>
	<i>Standard</i>	<i>New</i>	
Mouse	8	4	12
Track pad	1	9	10
Total	9	13	22

under the assumption that there is no association between the variables. Since 9 of the 22 participants (0.409) preferred the standard method, then if there were no association between pointing device and preferred method, one would expect 0.409 of the participants in the both the mouse condition and track pad condition to prefer the standard method. Of the 12 participants in the mouse condition, one would therefore expect $(0.409)(12) = 4.91$ participants to prefer the standard method. Similarly, in the track pad condition, one would expect $(0.409)(10) = 4.09$ participants to prefer the standard method. Note that this expected frequency is a mathematical concept; the number of participants in a sample with a specific preference would be a whole number.

An easy way to compute the expected frequency for a cell is to multiply the row total for the cell by the column total and then divide by the grand total. For the cell representing preference for the standard method when using the mouse, the expected frequency is $(12)(9)/22 = 4.91$. Table 2 shows the expected frequencies in parentheses.

The next step is to subtract, for each cell, the observed frequency from the expected frequency, square the difference, and then divide by the expected

Table 2 Data From the Example Experiment With Expected Frequencies

<i>Device</i>	<i>Preferred Method</i>		<i>Total</i>
	<i>Standard</i>	<i>New</i>	
Mouse	8 (4.91)	4 (7.09)	12
Track pad	1 (4.09)	9 (5.91)	10
Total	9	13	22

frequency. For the first cell, this is equal to $(4.91 - 8.00)^2/4.91 = 1.94$. The chi-square statistic is then computed by summing the values for all the cells. The formula can be written as

$$\chi^2 = \sum \frac{(E - O)^2}{E},$$

where E is an expected cell frequency and O is an observed cell frequency. For this example,

$$\begin{aligned} \chi^2 = & \frac{(4.91 - 8)^2}{4.91} + \frac{(7.09 - 4)^2}{7.09} \\ & + \frac{(4.09 - 1)^2}{4.09} + \frac{(5.91 - 9)^2}{5.91}, \end{aligned}$$

which is equal to 7.25. The chi-square statistic has a degrees of freedom parameter associated with it that is calculated as $(r - 1)(c - 1)$, where r is the number of rows in the data table and c is the number of columns. For this example, both r and c are equal to 2, so the degrees of freedom is $(2 - 1)(2 - 1) = 1$. The probability associated with a chi-square of 7.25 with one degree of freedom is 0.007. Since this is lower than conventional levels of significance, the null hypothesis of no association between preference and pointing device can be rejected, justifying the conclusion that there is an association in the population.

Assumptions and Accuracy of the Chi-Square Test

A key assumption of the test is that each observation is independent of each other observation. In general, this assumption is met if each participant in the experiment adds 1 to the frequency count of only one cell of the experiment. If this is the case, then the total frequency count will equal the total number of participants. However, even this safeguard does not ensure independence. For example, if 2 participants discussed their possible responses before making them, their responses would not be independent.

Suppose the example experiment had been a little different, with each participant making two preference

judgments, one when using the mouse and once when using the track pad. Since judgments from the same participant are not independent, the chi-square test of independence would not be appropriate for this design.

The statistic computed in the chi-square test of independence is only approximately distributed as chi-square, and therefore, the test is only an approximate test. Fortunately, under most circumstances, the approximation is quite good. Generally speaking, if there are at least 20 observations, the approximation will be satisfactory. However, accuracy is reduced if the proportions in the population are extreme. Suppose that the population proportion preferring the new method were 0.95 in both the mouse and track pad conditions. Since there is no association in the population, the null hypothesis is true. If the 0.05 significance level were used, the probability of making a Type I error should be 0.05. However, in this situation (with 12 and 10 participants, as in the experiment), the probability of a Type I error is 0.008. Thus, in this situation, the test is conservative: The true Type I error rate is lower than the nominal rate. In some situations, the actual Type I error rate is slightly higher than the nominal rate. For example, if the population proportions were both 0.60, then the probability of a Type I error would be 0.059 when the 0.05 significance level is used. Interestingly, for the same proportions, the probability of a Type I error using the 0.01 level is very close to 0.01. A simulation that allows one to estimate the probabilities of a Type I error in these designs is available at http://www.ruf.rice.edu/~lane/stat_sim/contingency/index.html.

Some writers have claimed that the chi-square test of independence assumes that all expected frequencies are greater than 5 and that a correction should be done otherwise. The correction is not recommended since, in general, it makes an already conservative test even more conservative.

Comparison to the Fisher Exact Probability Test

Since the chi-square test for independence is an approximate test, it would seem that the Fisher exact

probability test, which, as its name implies, results in exact probabilities, would be preferable. However, this is not generally the case. The Fisher exact probability test is based on the assumption that the row and column totals are known before the experiment is conducted. For example, consider a hypothetical experiment designed to determine whether caffeine decreases response time. Suppose 10 participants are given caffeine and an additional 10 participants are given a placebo. After response times are measured, every participant below the median is classified as a fast responder, and every participant above the median is classified as a slow responder. Therefore, even before the data are collected, it is known that the column totals will be as shown in Table 3.

By contrast, in the example experiment, it was not known until the data were collected that 9 participants would prefer the standard method and 13 would prefer the new method. When the Fisher exact probability test is used with designs in which the row and column totals are not known in advance, it is very conservative, resulting in Type I error rates below the nominal significance level.

Using the Computer

For this example, the statistical package SAS JMP was used to compute the test. The output in Figure 1 shows the proportion preferring the new and the standard methods as a function of device. JMP automatically reports the results of two methods of testing for significance. The Pearson chi-square is the method discussed here. As can be seen in Table 4, the chi-square of 7.246 and probability of 0.0071 match the results computed previously.

Table 3 Row and Column Totals in Caffeine Study

<i>Condition</i>			
<i>Speed</i>	<i>Drug</i>	<i>Placebo</i>	<i>Total</i>
Fast			10
Slow			10
Total	10	10	22

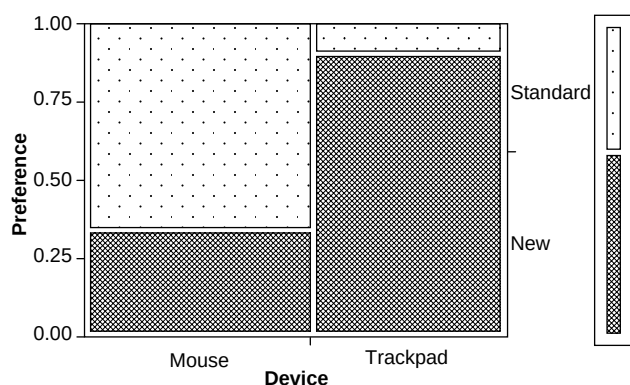


Figure 1 Proportion Preferring the New and the Standard Methods as a Function of Device

Table 4 Text Output From JMP

Test	Chi-square	Prob > chi sq
Likelihood ratio	7.989	0.0047
Pearson	7.246	0.0071

Note: Prob = probability; chi sq = chi-square.

Larger Designs

The example experiment is a 2×2 design since the two variables (device and method) each have two levels. The chi-square test for independence can also be used for larger designs. For instance, the example experiment might have used three devices and four methods, resulting in a 3×4 design. The calculations for doing the test are the same except that there are more cells over which the values of $(E - O)^2/E$ are summed. The degrees of freedom would be $(3 - 1)(4 - 1) = 6$.

—David M. Lane

Further Reading

- Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi square test of independence in $r \times c$ tables that have small expected frequencies. *Psychological Bulletin*, 86, 1290–1297.
- Lane, D. M., Peres, S. C., Sándor, A., & Napier, H. A. (2005). A process for anticipating and executing icon selection in

graphical user interfaces. *International Journal of Human Computer Interaction*, 19, 241–252.

Java applet to compute probability values for chi-square:
http://psych.rice.edu/online_stat/analysis_lab/chi_square_prob.html

Java applet simulation of the chi-square test of association:
http://www.ruf.rice.edu/~lane/stat_sim/contingency/index.html

CHILDREN'S ACADEMIC INTRINSIC MOTIVATION INVENTORY

The Children's Academic Intrinsic Motivation Inventory (CAIMI), published by Psychological Assessment Resources (www.parinc.com), measures intrinsic motivation for school learning. CAIMI items are based on theories of intrinsic motivation measuring enjoyment of learning; an orientation toward mastery; curiosity; persistence; and the learning of challenging, difficult, and novel tasks. It is a self-report instrument consisting of 44 items, to which children rate their agreement or disagreement. There are five subscales, four being subject-area specific (reading, math, social studies, and science) and one addressing school in general. The CAIMI was originally developed (in 1985) for students in Grades 4 through 9 and in 2001 was extended through the end of high school. A modified downward extension (YCAIMI) for Grades 1 through 3 was developed in 1990.

The CAIMI may be administered to a group or to individuals and in a classroom or an office setting. Children with sufficient reading ability may complete the CAIMI on their own after instructions and practice items are read aloud. Individual, oral administration is recommended for those with learning, reading, or perceptual difficulties. Individual administration takes approximately 20–30 minutes, and for administration to a group, sufficient time must be allocated for distribution and collection of materials, bringing total time to about an hour. Percentiles and *T* scores are available. Advantages of the CAIMI are that it allows for distinguishing motivation from ability and

achievement, provides a motivational profile across the four subject areas and for school in general, is easily administered and scored, and may be applied to a variety of settings. For example, the CAIMI can be used by psychologists and practitioners; teachers and school administrators in regular and special education, including programs for gifted students; program evaluators; and researchers. It has been used by school districts to help identify children for inclusion in programs for the gifted.

The CAIMI has excellent psychometric properties and has been used nationally and internationally. It has been translated into several languages, including Spanish, Japanese, Chinese, and Slovene. Major research findings made with the CAIMI include the following: (a) Motivation is uniquely related to achievement above and beyond IQ; (b) the CAIMI provides stable measurement of children's motivation from upper elementary school through the end of high school; (c) children with higher academic intrinsic motivation function more effectively in school (higher achievement, more positive self-perception of performance, lower academic anxiety, lower extrinsic motivation) from the elementary school years through the end of high school; (d) children whose parents encourage intrinsic motivation and provide a stimulating environment have greater academic intrinsic motivation; (e) intellectually and motivationally gifted children have significantly higher academic intrinsic motivation; and (f) children with exceptionally low motivation (motivationally disadvantaged children) can be identified as early as Grade 4, and such children evidence a variety of associated poor school functioning from that time through the end of high school. Research has been conducted using the CAIMI in the Fullerton Longitudinal Study, in school settings, and with a variety of children, including students in regular and special education populations. Thus, the validity of the CAIMI has been generalized across a variety of populations and settings.

—Adele Eskeles Gottfried

See also Achievement Tests

Further Reading

- Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology*, 77, 631–635.
- Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology*, 82, 525–538.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (1994). Role of parental motivational practices in children's academic intrinsic motivation and achievement. *Journal of Educational Psychology*, 86, 104–113.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (1998). Role of cognitively stimulating home environment in children's academic intrinsic motivation: A longitudinal study. *Child Development*, 69, 1448–1460.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, 93, 3–13.
- Gottfried, A. W., Gottfried, A. E., Cook, C., & Morris, P. (2005). Educational characteristics of adolescents with gifted academic intrinsic motivation: A longitudinal study from school entry through early adulthood. *Gifted Child Quarterly*, 49, 172–186.

CLASS INTERVAL

A fundamental skill in measurement and statistics (and all of science, for that matter) is being able to communicate the greatest amount of information about a data set using as little effort as possible. Such is the case when it comes to consolidating a data set to represent it by a frequency distribution. The first part in the creation of a frequency distribution is the establishment of a class interval, or the range of values that constitutes a category. A class interval is a range of numbers. Class intervals are also called *bins* (something you put data in), *class boundaries*, and *class limits*.

The first step in establishing a class interval is defining how large each class interval will be. In the frequency distribution in Figure 1 (based on the data set that is also shown), each interval spans five possible scores, such as 5–9 (which includes scores 5, 6, 7, 8, and 9) and 40–44 (which includes scores 40, 41, 42, 43, and 44).

The raw data . . .

47	10	31	25	20
2	11	31	25	21
44	14	15	26	21
41	14	16	26	21
7	30	17	27	24
6	30	16	29	24
35	32	15	29	23
38	33	19	28	20
35	34	18	29	21
36	32	16	27	20

The frequency distribution . . .

<i>Class Interval</i>	<i>Frequency</i>
45–49	1
40–44	2
35–39	4
30–34	8
25–29	10
20–24	10
15–19	8
10–14	4
5–9	2
0–4	1

Figure 1 Frequency Distribution

Here are some general rules to follow in the creation of a class interval, regardless of the size of values in the data set you are dealing with.

1. Select a class interval that has a range of 2, 5, 10, or 20 data points. In the example in Figure 1, a class interval of 5 was used.
2. Select a class interval so that 10 to 20 such intervals cover the entire range of data. A convenient way to do this is to compute the range, then divide by a number that represents the number of intervals you want to use (between 10 and 20). In this example, there are 50 scores, and 10 intervals are being used, so $50/10 = 5$, which is the size of each class interval. For example, if scores ranged from 100 to 400, $300/20 = 15$, so 15 would be the class interval.
3. Begin listing the class interval with a multiple of that interval. In the sample frequency distribution, the class interval is 5 and it started with the lowest class interval of 0.
4. Finally, place the largest interval at the top of the frequency distribution.

Once class intervals are created, the frequency part of the frequency distribution can be completed. That includes counting the number of times a score occurs in the raw data and entering that number in each of the class intervals represented by the count.

—Neil J. Salkind

See also Frequency Distribution

Further Reading

Pfeiffer, C., Windzio, M., & Kleimann, M. (2005). Media use and its impacts on crime perception, sentencing attitudes and crime policy. *European Journal of Criminology*, 2(3), 259–285.

Class interval discussion: Scale and impression: <http://www.shodor.org/interactivate/discussions/sd2.html>

CLASSICAL TEST THEORY

The first theory of measurement has been named classical test theory (CTT) because it was formulated from simple assumptions made by test theorists at the start of testing. It is also called the *theory of true and error scores*, because it is thought to comprise both true scores and error, and *classical reliability theory*, since its major function is to evaluate the reliability of the observed scores on a test; that is, it calculates the strength of the relationship between the observed score and the true score.

CTT makes use of a number of models considered as a group and various procedures with which the test developer tries to provide solutions to complicated problems in order to measure psychological variables. Psychologists are not interested in the score itself but in the conclusions they can draw and the explanations they can make on the basis of the measured behavior of an individual.

Assumptions of CTT

CTT views a test score as having two components, the *true score* and a *random error*. The true score is considered the average of identical values taken from repeated measurements without limit. The error is regarded as unrelated either to the true score or to the error that can occur in another measurement of the same attribute. Although the theory is an oversimplification and does not represent the results, it brings out relationships that are informative and useful in test design and construction as well as in evaluation of test scores. The first basic assumption of CTT is that the obtained score is the sum of true score plus error; that is, the true score and the error score are inextricably mixed. This concept can be expressed as a simple equation:

$$X = T + E,$$

where

X is the obtained score,

T is the true score, and

E represents errors of measurement.

It must be pointed out that the true score is never known. It remains within a certain interval, however, and a best estimate of it can be obtained.

Measurement error is all things except the true score. Errors of measurement can arise from numerous sources, such as item selection, test administration, test scoring, and systematic errors of measurement. The first three sources of error are jointly called *unsystematic measurement error*, meaning that their impact is unexpected and inconsistent. A systematic measurement error occurs when a test consistently measures something different from the trait it was designed to measure.

Measurement error reduces reliability or repeatability of test results. The assumption that the obtained score is made up of the true score and the error score reveals several additional assumptions. An assumption derived from true scores is that unsystematic measurement error affects test scores randomly. The randomness of measurement error is a fundamental assumption of CTT. Since there are random events,

unsystematic measurement errors have some probability of being positive or negative, and consequently they amount to an average of zero across a large group of subjects. It follows that the mean error of measurement is zero. Another assumption of CTT is that measurement errors are not correlated with true scores. A final assumption is that measurement errors are not correlated with errors on other tests. All these assumptions can be summarized as follows: (a) Measurement errors are random, (b) the mean error of measurement is zero, (c) true scores and error are uncorrelated, and (d) errors on different tests are uncorrelated.

From the aforementioned assumptions, we can arrive at some significant conclusions about the reliability of measurement; for example, when we administer a test to a large number of persons, we find variability of scores that can be expressed as a variance, σ^2 . According to CTT, the variance of obtained scores has two separate sources, the variance of the true score and the variance of errors of measurement, or

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2,$$

where

σ_X^2 is the variance of obtained scores,

σ_T^2 is the variance of true scores, and

σ_E^2 is the variance errors of measurement.

The second basic assumption of CTT is that of parallel tests. The term *parallel tests* comes from the fact that each individual item can be viewed as a test; that is, each item coincides with the value of the latent variable. The assumption of the parallel tests model is satisfactory because it leads to useful conclusions about the relationships of the individual items to the latent variable that are grounded in observations of the relations of the items to one another. Thus, the parallel tests model adds two assumptions to CTT:

1. The amount that the latent variable affects each item is regarded as the same for all items.
2. Each item is considered to have the same amount of error as any other item. This means that the factors affect all items equally.

These assumptions imply that the correlation of each item with the true score is identical, and this conclusion leads to quantifying reliability.

In summary, the parallel tests model presupposes: (a) random error, (b) errors that are uncorrelated with each other, (c) errors that are uncorrelated with the true score, (d) the latent variable's influencing all items equally, and (e) an equal amount of error for every item. Thus the model allows us to make inferences about the latent variable that are grounded in the correlation between the items. However, the model achieves this by proposing strict assumptions.

Alternative Models to the Parallel Model

Apart from the strict assumptions referred to above, useful inferences can be made with less-rigid assumptions. Another model is based on what is called *essentially tau-equivalent tests* and makes a more tolerant assumption, namely, that each item requires the same amount of error variance that other items have. This means that the items are parallel with regard to the latent variable but are not essentially affected to the same degree by extraneous factors that are called *error*. Thus, since error can vary, item means and item variances can also vary. This model allows us to reach the same conclusion as the parallel tests model does, but with more lenient assumptions.

However, some theorists think that even the essentially tau-equivalent model is too rigid. For this reason, they put forward what is called the *congeneric model*, with even more liberal assumptions. The model assumes only that all items have a common latent variable. The items do not have equally strong relationships to the latent variable, and their error variances are not required to be equal. The only thing that we must assume is that each item exhibits the true score to some degree.

A Brief History of CTT

In 1904, Charles Spearman put forward CTT, and it is still flourishing. Spearman developed CTT by combining the concept of error with the concept of

correlation. He argued that test scores have an error when we measure a human trait, and thus the observed correlation between test scores is lower than the correlation between true scores. Later, other authors restated and worked out the theory. Guilford, for example, tried to synthesize the developments that had taken place in CTT through 1936, Gulliksen gave a detailed account of all the progress through 1950 in his classic work, *Theory of Mental Tests*, and Lord and Novick reanalyzed CTT and combined it with new psychometric theories in 1968. However, the early days of the theory were rather difficult because of the quantitative atmosphere prevalent at that time, and its acceptance was lukewarm at first. However, before long, tests were constructed based on CTT, psychometric technology grew at a rapid rate, significant progress was achieved in psychological assessment, and thus the theory was extended and stabilized.

Controversies Surrounding CTT and Its Impact

CTT is criticized because of its simplicity and because the concept of true score is only a notion. However, new theories, such as item response theory (IRT) and generalizability theory supplement CTT and can never replace it. Most psychological tests currently available are constructed according to CTT. This means that new theories have to be combined with CTT in order to surmount the problems associated with it.

CTT is characterized by its simplicity and flexibility and for this reason can be used in many different circumstances. These qualities guarantee that CTT will be used over and over again in the future. Moreover, the new psychometric models that comprise item response theory, although they have developed rapidly in the past 35 years, have also demonstrated unfulfilled assumptions and uninterpretable parameters, as well as estimation difficulties. For these reasons, the two- and three-parameter models may become less popular in the future. This may lead to revival of interest in CTT, and its use and practice may extend further. Furthermore, the availability and development of computers will make calculating

true score, establishing norms, and equating tests easier, no doubt increasing the use of CTT even further.

—Demetrios S. Alexopoulos

See also Item Response Theory; Reliability Theory; Validity Theory

Further Reading

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). New York: McGraw-Hill.
- Spearman, C. E. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 205–293.

CLASSIFICATION AND REGRESSION TREE

Classification and regression tree (CART) is a machine learning (or classification) algorithm that constructs a tree-structured classifier to assign group labels to each case based on its attributes. The resulting tree-structured classifier is usually ideal for interpretation and decision making. The algorithm requires that for each case in the data, there be two variables. The first is the group variable to be classified and predicted (such as disease status or treatment), and the second is the variable of attributes that can be multi-dimensional numerical or categorical data (such as smoking status, sex, or abundance of various enzymes in blood). Normally, the method is implemented in a set of training cases to learn a classifier. The classifier is then applied to an independent test set to evaluate the generalized classification accuracy.

CART analysis is performed as a binary recursive partition tree. An impurity measure is defined to describe the purity (concentration in a single group) of cases in a node. The algorithm recursively searches for

the attribute criterion that partitions data into two parts with the largest decrease of impurity measure. Normally, the Gini impurity index is used in CART analysis:

$$I(T) = 1 - \sum_{i=1}^m p_i^2,$$

where

T is the data set in the node,

m the number of groups, and

p_i the proportion of group i in the node.

When all cases in the node belong to the same group (the purest case), $I(T)$ is minimized at 0, and when each group has the same proportion (the most impure case), $I(T)$ is maximized. The partition of the hierarchical tree continues until either the number of cases in the node is too small or the decrease of the impurity index is not statistically significant. Additional pruning rules may be applied to decide termination of tree growth to prevent a problem of overfitting.

CART has a number of merits compared with other classification algorithms. The method is inherently nonparametric without a distribution assumption of the data (in contrast to methods like linear discriminant analysis). It is thus more robust against skewed or ill-behaved distributed data. Learning in CART is a “white” box, and the learned classification criteria are easy to interpret. Thus CART is most applicable in situations when interpretation and learning of important attributes contributing to classification are the major goals in the application. CART by its nature can easily handle categorical and ordinal data in addition to numerical data. Finally, computation of an exhaustive search for the best partition is very fast, making the method feasible for large data sets.

Software Package

Since CART is a relatively modern statistical technique, it is not implemented in most major statistical software (e.g., SAS and S-PLUS). SPSS contains an add-on module, “SPSS Classification Trees.” A commercial software, “Classification and Regression Tree,” specifically for CART analysis, is also available. In the following discussion, an extension package “tree” of the free software R is used to implement CART.

Table 1 MPG for a Select Group of Cars

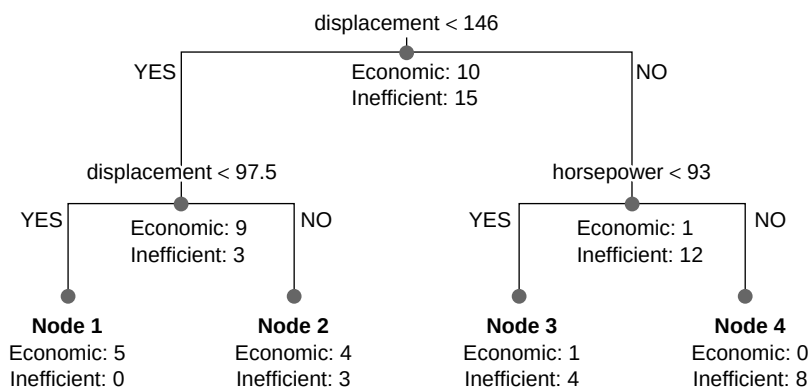
<i>Car</i>	<i>Efficiency</i>	<i>Cylinders</i>	<i>Displacement</i>	<i>Horsepower</i>	<i>Weight</i>	<i>Acceleration</i>
1	Inefficient	4	151	85	2855	17.6
2	Economic	4	98	76	2144	14.7
3	Economic	5	121	67	2950	19.9
4	Inefficient	6	250	105	3353	14.5
5	Inefficient	4	151	88	2740	16
6	Inefficient	6	250	88	3021	16.5
7	Economic	4	71	65	1836	21
8	Economic	4	112	88	2395	18
9	Economic	4	141	71	3190	24.8
10	Inefficient	8	350	155	4360	14.9
11	Inefficient	4	98	60	2164	22.1
12	Economic	6	262	85	3015	17
13	Inefficient	6	200	85	3070	16.7
14	Inefficient	6	258	110	2962	13.5
15	Inefficient	4	116	75	2158	15.5
16	Inefficient	4	140	72	2401	19.5
17	Inefficient	8	350	180	4499	12.5
18	Inefficient	8	307	200	4376	15
19	Inefficient	8	318	140	3735	13.2
20	Economic	4	78	52	1985	19.4
21	Economic	4	89	71	1990	14.9
22	Economic	4	97	75	2265	18.2
23	Inefficient	6	250	98	3525	19
24	Economic	4	83	61	2003	19
25	Inefficient	8	302	140	3449	10.5

An Example

An example of classification of car fuel efficiency is demonstrated as follows. The data shown in Table 1 are a random subsample of 25 cars from the “auto-mpg” data set from the UCI Machine Learning Repository (<http://www.ics.uci.edu/~mlearn/MLRepository.html>). The group variable for classification is “efficiency,” which has two possible values: inefficient (mpg < 25) and economic (mpg ≥ 25). Five attributes for classifying and predicting fuel efficiency are available for prediction: cylinders, displacement, horsepower, weight, and acceleration.

The CART output of this example is shown in Figure 1. The top node

represents the whole data set, containing 10 economic and 15 inefficient cars. The algorithm finds that the best way to classify car fuel efficiency is “whether the displacement is larger than

**Figure 1** CART Output of Classification of Car Fuel Efficiency

146 or not.” This will split the data into two groups. The first (“YES”) group (displacement < 146) contains 9 economic and 3 inefficient cars, and the second (“NO”) group (displacement > 146) contains 1 economic and 12 inefficient cars. Similarly, the “displacement < 146” group is best split into Node 1 (displacement < 97.5) and Node 2 (97.5 < displacement < 146). The “displacement > 146” group is split by horsepower to generate Node 3 (displacement > 146 & horsepower < 93) and Node 4 (displacement > 146 & horsepower > 93). All nodes except for Node 2 have high classification accuracy in this training data. It should be noted that attributes used for classification may appear repeatedly in different branches (e.g., displacement in this example). Figure 2 demonstrates a scatter plot of the 25 cars. The solid lines represent the splitting rules in the branches of the CART output.

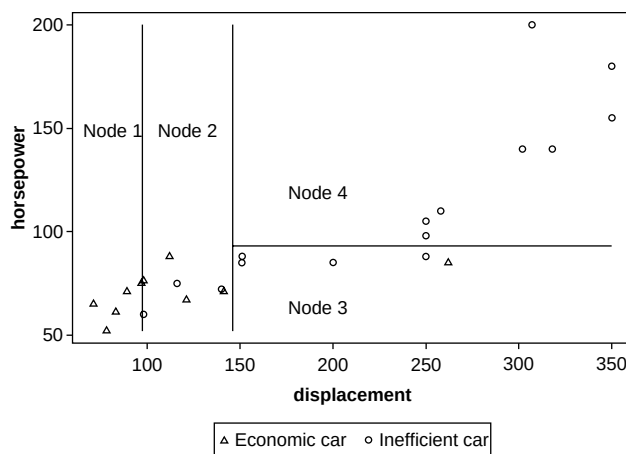


Figure 2 Scatter Plot of Subsample of 25 Cars

—George C. Tseng

Further Reading

Breiman, L. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman & Hall/CRC.

Classification and regression tree downloads: <http://cran.au.r-project.org/src/contrib/Descriptions/tree.html>

CLINICAL ASSESSMENT OF ATTENTION DEFICIT

The Clinical Assessment of Attention Deficit (CAT) consists of two comprehensive assessment tools for the objective evaluation of attention deficit disorder and attention deficit/hyperactivity disorder (ADD/ADHD) symptoms in children, adolescents, and adults between 8 and 79 years of age. The CAT is especially useful in clinical, educational, and research settings because it is theoretically conceived and psychometrically sound. Although the CAT is very comprehensive in the context-related behaviors and experiences it assesses, it is brief (10–20 minutes) and easily administered. The CAT can be (a) administered individually or in groups, (b) mailed to respondents' homes to be completed and returned, or (c) completed in its paper-and-pencil format in a professional's office. The CAT Scoring Program performs all scoring and profiling of the assessment and provides a very thorough interpretative report.

The CAT includes a 108-item instrument for adults (CAT-A) and a 42-item instrument for children and adolescents (CAT-C). These instruments are very similar in structure, format, and item content. Each instrument assesses important clinical behaviors related to ADD with and without hyperactivity (as outlined in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., text rev.) via three clinical scales (inattention, impulsivity, and hyperactivity). Both instruments consider these important clinical behaviors within multiple contexts (as suggested by the American Academy of Pediatrics) by means of the CAT context scales (personal, academic/occupational, and social) and within an individual's personal experiences through the CAT locus scales (internal and external). Figure 1 depicts the CAT theoretical model and shows the clinical, context, and locus scales. The CAT-C and CAT-A assess the same clinical behaviors within the same contexts and consider the same internal feelings and external behaviors.

Context Scales Locus Scales		Clinical Subscales		
<i>Personal</i>		<i>Inattention</i>	<i>Impulsivity</i>	<i>Hyperactivity</i>
Internal	I was bored easily.	I was impatient.	I often felt nervous.	
External	It seems I was always spilling something.	I acted without thinking.	I was extremely active.	
<i>Academic Occupational</i>				
Internal	I daydreamed a lot in school.	I often regretted my actions in school.	I felt very restless in school.	
External	I was frequently tardy for class.	I blurted out answers to questions in class.	I talked too much in school.	
<i>Social</i>				
Internal	I often could not recall the names of people I met.	I made quick judgments about other people.	I always had more energy than my friends.	
External	When playing games, I often played out of turn.	I took more risks than my friends.	I rarely played quietly with my friends.	

Figure 1 CAT-A Conceptual Blueprint (Sample Items From the Childhood Memories Scale)

The minor differences between the two instruments are primarily age-related considerations. Because adult diagnosis of ADD/ADHD requires evidence of both childhood onset and current symptoms, the CAT-A is composed of two subparts: Part I, Current Symptoms, and Part II, Childhood Memories. In contrast, CAT-C assesses only current ADD/ADHD symptoms and therefore is not divided into age-related subparts. Also, CAT-A includes only a self-report form due to the limited availability of raters who could accurately recall an adult client's childhood behaviors. CAT-C includes forms for self, parent, and teacher because of readily available respondents who can reliably report on a child's current behaviors across multiple settings.

CAT normative samples and total scale (Clinical Index) coefficient alpha reliabilities are quite consistent across instruments and forms. The CAT-C self form was normed on 800 children and adolescents 8 to 18 years old; the CAT-C parent form was normed using the ratings from the parents of the 800 children and adolescents who completed the self form. The CAT-C teacher form also was normed on 550 teachers of the same children and adolescents who completed

the self form. CAT-A was normed on 800 adults ranging from 19 to 79 years old. CAT-C Clinical Index internal consistency coefficients range from .92 on the self form to .98 on the teacher form. The CAT-A Clinical Index is .94 for Childhood Memories and .91 for Current Symptoms, with an overall CAT-A Clinical Index (i.e., combining Childhood Memories and Current Symptoms) of .96.

—Bruce A. Bracken

Further Reading

- American Academy of Pediatrics. (2000). Clinical practice guideline: Diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. *Pediatrics*, 105I, 1158–1170.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Bracken, B. A., & Boatwright, B. S. (2005). *Clinical assessment of attention deficit—adult*. Lutz, FL: Psychological Assessment Resources.
- Bracken, B. A., & Boatwright, B. S. (2005). *Clinical assessment of attention deficit—child*. Lutz, FL: Psychological Assessment Resources.

CLINICAL ASSESSMENT OF BEHAVIOR

The Clinical Assessment of Behavior (CAB) is a comprehensive, third-party rating scale for the behavioral assessment of children and adolescents ages 2 to 18 years. The two parent forms (a 170-item CAB—Parent Extended form and a 70-item CAB—Parent form) and the 70-item teacher form (CAB—Teacher) allow for a balanced assessment of both adaptive and clinical (i.e., maladaptive) behaviors across various settings, contexts, and raters. With a single set of items spanning the age range of 2 to 18 years, the CAB permits a longitudinal evaluation of symptoms throughout the course of an individual's treatment, without the introduction of potentially incompatible instruments into the evaluation process as the individual ages.

The CAB structural content includes 3 adaptive and 3 clinical scales and 2 adaptive and 10 clinical clusters. Whereas the CAB scales were designed to reflect "broad brush" psychosocial adjustment domains, CAB clusters reflect specific areas of exceptionality or disorder. Table 1 illustrates the exceptionalities, disorders, and conditions assessed by the CAB scales and clusters across the three forms. Table 1 also presents coefficient alpha reliability indices for each scale and cluster for the standardization sample. With its close alignment to the diagnostic criteria of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.), legislative mandates of the Individuals with Disabilities Education Act, and a context-dependent model of psychosocial adjustment, the CAB augments the diagnosis of childhood and adolescent psychiatric disorders with comprehensive behavioral content.

The CAB forms require an eighth-grade reading level of respondents and 5–10 minutes to complete the CAB-P and CAB-T forms (the CAB-PX requires 10–20 minutes). Because it is brief and easy to read, the CAB can be completed by raters without the aid of psychologists, thus rendering the evaluation process more efficient than if parents and teachers were to be interviewed. Furthermore, the CAB is computer scored and interpreted, and the resulting data are profiled and reported for professional interpretation and use.

A relatively unique application of the CAB is the ability for professionals to use its data to identify specific behaviors associated with educationally relevant exceptionalities (e.g., mental retardation, learning disabilities, giftedness and talent, attention deficit disorder, and attention deficit/hyperactivity disorder).

Table 1 CAB Scales, Clusters, and Coefficient Alpha Reliabilities for the CAB-PX, CAB-P, and CAB-T

Scale/Cluster	CAB-PX	CAB-P	CAB-T
Clinical scales			
Internalizing behaviors	.95	.89	.92
Externalizing behaviors	.97	.95	.98
Critical behaviors	.91	—	—
Adaptive scales			
Social skills	.95	.92	.96
Competence	.94	.92	.96
Adaptive behaviors	.92	—	—
Clinical clusters			
Anxiety	.93	.88	.92
Depression	.95	.90	.93
Anger	.93	.90	.94
Aggression	.95	.92	.97
Bullying	.97	.94	.97
Conduct problems	.92	.90	.96
Attention deficit/hyperactivity	.94	.94	.97
Autistic spectrum disorders	.92	.89	.93
Learning disability	.92	.90	.95
Mental retardation	.91	.90	.95
Adaptive clusters			
Executive function	.91	.91	.95
Gifted and talented	.94	.92	.96
CAB Behavioral Index	.98	.97	.99

Notes: CAB = Clinical Assessment of Behavior; PX = Parent Extended form; P = Parent form; T = Teacher form.

Too often, educational exceptionalities are diagnosed with only ability measures (e.g., intelligence tests, achievement tests, perceptual/motor tests) and without questioning whether a student's behaviors are consistent with the considered diagnosis. The CAB allows third-party respondents to indicate the frequency with which they observe the child or adolescent demonstrating behavioral characteristics associated with specific educational exceptionalities. Such information provides a multisource, multicontext way to corroborate referral information and ability test data with behavioral indices.

Standardized on more than 2,100 parents and 1,600 teachers, the CAB normative sample is highly representative of the U.S. population. An examination of the extent to which students' behavioral functioning is associated with their demographic characteristics (age, gender, race/ethnicity, parents' educational levels) revealed that generally 3% or less of the variance in the CAB parent-generated ratings was associated with the demographic attributes evaluated. Approximately 9% or less of the variance in teachers' ratings was associated with demographic attributes. These empirical findings, along with comparable reliability coefficients for all age, gender, and ethnic/racial groups, suggest that the CAB may provide the basis for an equitable behavioral assessment regardless of a student's age, gender, race/ethnicity, or socioeconomic status.

—Bruce A. Bracken

Further Reading

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Bracken, B. A. (1992). *Multidimensional self concept scale*. Austin, TX: PRO-ED.
- Bracken, B. A. (1996). Clinical applications of a multidimensional, context-dependent model of self-concept. In B. A. Bracken (Ed.), *Handbook of self concept: Developmental, social, and clinical considerations* (pp. 463–505). New York: Wiley.
- Bracken, B. A., & Keith, L. K. (2004). *Clinical assessment of behavior*. Lutz, FL: Psychological Assessment Resources.
- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 103-218 (GPO 1997).

CLINICAL ASSESSMENT OF DEPRESSION

The Clinical Assessment of Depression (CAD) is a comprehensive assessment of children's, adolescents', and adults' depressive symptoms. CAD content was developed from a review of the literature pertaining to child, adolescent, and adult development and depression and was closely aligned with current diagnostic criteria of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.).

Notably, the CAD employs a single 50-item form as an overall measure of general affectivity. This comprehensive set of depressive symptoms was employed in part to test previous assumptions that the nature of depression varies across the age span. Multidimensionality was added to the instrument through the inclusion of symptom scales and critical item clusters sensitive to specific dimensions of depression. The CAD produces a total scale score and standard scores for each of four symptom scales and six clusters. The CAD critical item clusters (e.g., Hopelessness, Self-Devaluation) include item content that is especially sensitive to individuals who may be at risk for harming themselves. Table 1 shows the full range of affective symptoms assessed on the CAD and its symptom scales.

The CAD was normed on a sample of 1,900 children, adolescents, and adults aged 8 through 79 years. The normative sample included individuals from a wide range of racial and ethnic backgrounds, geographical

Table 1 Coefficient Alpha Reliabilities for the CAD Symptom Scales by Age

	Age in Years			
	8–11	12–17	18–25	26–79
Depressed mood	.95	.96	.96	.95
Anxiety/worry	.83	.85	.83	.86
Diminished interest	.78	.85	.85	.86
Cognitive and				
physical fatigue	.82	.83	.85	.87
CAD total scale	.96	.97	.96	.97

regions of the United States, and residential communities (urban, suburban, and rural). Table 1 also presents the instrument's total sample symptom scale and total scale internal consistency estimates; CAD total scale reliabilities range from .96 to .97 across the entire age range. Less than 1% of the total scale variability is associated with examinees' age, gender, race/ethnicity, or socioeconomic status. Overall, data from CAD reliability and validity studies suggest that depressive symptoms are common and behave similarly, regardless of the demographic characteristics of the examinee.

The CAD has multiple applications, with broad uses in clinical, educational, and research settings. Although it is very comprehensive as a measure of depressive symptomatology, the CAD is appropriately brief (10 minutes) for use with depressed clients and is easily administered. Because the CAD is a self-report instrument requiring only a fourth-grade reading level, it can be completed by most clients, regardless of age, without help from an examiner. Maximizing scoring and interpretation efficiency, the CAD can be scored either by hand or by computer by individuals with little psychometric training, and resulting data can be collated, profiled, and reported for professional interpretation and use.

—Bruce A. Bracken

Further Reading

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Bracken, B. A., & Howell, K. K. (2005). *Clinical assessment of depression*. Lutz, FL: Psychological Assessment Resources.
- Finch, S. M. (1960). *Fundamentals of child psychiatry*. New York: Norton.
- Rie, F. H. (1966). Depression in childhood: A survey of some pertinent contributions. *Journal of Child Psychology and Psychiatry and Applied Disciplines*, 35(7), 1289–1308.

CLUSTER ANALYSIS

The objective of cluster analysis is to construct a natural grouping of the objects in a multivariate data

set. The grouping captures “similarities” between the objects according to some criterion. An ideal outcome is one in which the objects belonging to the same group are as similar as possible, whereas objects belonging to different groups are as dissimilar as possible. Cluster analysis has proved to be a successful data reduction technique in many scientific fields.

The main steps in cluster analysis are (a) the choice of algorithms for constructing groups (clusters) and (b) the choice of a similarity-dissimilarity measure. Popular choices of dissimilarity measures are various distances, such as Euclidean or Manhattan. An interesting, robust distance measure is discussed in Kaufman and Rousseeuw.

Clustering Algorithms

Clustering algorithms can be classified into two groups: (a) *partition* methods and (b) *hierarchical* methods. Partition methods create a family of clusters in which each object belongs to just a single member of the partition. The requirement to generate such partitions is that distances between pairs of objects belonging to the same cluster are smaller than distances between pairs of objects in different clusters. Formally, suppose that objects i and j belong to cluster A , while object k belongs to cluster B . It is then required that $d_{ij} < d_{ik}$ and $d_{ij} < d_{jk}$, where d_{ij} denotes a dissimilarity measure for objects i and j . The most prominent partition algorithm is k - means. A more robust alternative (partitioning around medoids) is discussed in Kaufman and Rousseeuw, while a model-based version is presented in Banfield and Raftery.

Agglomerative hierarchical algorithms construct a clustering solution that is represented by a tree (dendrogram). Their characteristic feature is that either any two clusters are disjoint or one cluster is a superset of the other. It is usually required for the dissimilarity measure to be an ultrametric; in other words, for every triple set of objects (i, j, k) we have, the two largest values in the set $\{d_{ij}, d_{jk}, d_{ik}\}$ are equal. An agglomerative algorithm starts with as many clusters as objects in the data set and progressively merges them, based on the shortest distance between them, until each object has been assigned to a single cluster.

The mechanism used for merging clusters is based on a distance measure between groups. Some common possibilities include the shortest distance between members of the two groups (single linkage), the largest one (complete linkage), the average distance (average linkage), and the Euclidean distance between the averages of the clusters (centroid method). Some other choices are discussed in Gordon.

An important practical consideration is that since cluster analysis is driven by the distance between objects, variables need to be scaled appropriately so as not to exert excessive influence on the distance measure.

—George Michailidis

See also Factor Analysis

Further Reading

- Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London: Hodder Arnold.
- Gordon, A. D. (1999). *Classification*. London: Chapman & Hall.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: Wiley.
- Mirkin, B. (2005). *Clustering for data mining: A data recovery approach*. London: Chapman & Hall.

CLUSTER SAMPLING

A cluster sample is a type of sample generated for the purposes of describing a population in which the units, or elements, of the population are organized into groups, called clusters. The goal of a survey is to gather data in order to describe the characteristics of a population. A population can consist of individuals, school districts, plots of land, or a company's invoices. A survey collects information on a sample, or subset, of the population. In cluster sampling, instead of choosing individuals to interview or units on which to record data directly, the survey researcher first chooses clusters of elements. Once

clusters are selected, one often then samples elements from within the selected clusters and collects data on them. A survey following this procedure with two stages of random selection of units is called a *two-stage cluster sample*. A survey that selects units using three or more stages of random selection is called a three-stage or a multistage cluster sample. The groups of units selected at the first stage of sampling are called *primary sampling units* (PSUs). The groups selected at the second stage are called *secondary sampling units* (SSUs). The elements selected at the final stage can be referred to as the *ultimate sampling units*.

Cluster sampling is used widely in large-scale official surveys. The U.S. Department of Agriculture's National Resources Inventory study selects land areas, called segments, as PSUs. It then selects points, defined by latitude and longitude, as SSUs. The U.S. Bureau of Labor's Current Population Survey, which is conducted by the U.S. Bureau of the Census, selects counties as PSUs and then households within counties. Surveys of school students organized by the National Center for Education Statistics and surveys focused on health and nutrition designed by the National Center for Health Statistics rely on sampling designs that involve clustering.

Cluster sampling can be a type of probability sampling, which means that it is possible to compute the probability of selecting any particular sample. The main benefit of probability sampling is that one can estimate means, proportions, and variances without the problem of selection bias. In comparison with *simple random sampling* (SRS), in which units are selected directly from the population list or frame, estimates from a cluster sample design usually have greater variability or random uncertainty. The increase in variance of estimators occurs because each initial selection of a cluster means that multiple units ultimately will be in the sample. Units clustered together typically are more alike than randomly selected units from across the entire population. As a result, each initial selection under cluster sampling has more impact on resulting estimates than it does under SRS. If clusters are small and if two or more stages of random selection are used, the impact of cluster sampling

on variances is not too large. The main advantage of cluster sampling is convenience and reduced cost. If one wanted to interview high school students or look at their transcripts in the state of Iowa, an SRS would require sending data collectors all over the state and negotiating permissions with numerous school districts to interview students and look at confidential data. A cluster sample could first select school districts and then schools within districts before selecting students. Fewer schools would need to be visited, thereby reducing travel and setup costs and time. Although cluster sampling is convenient, it is not the same thing as *convenience sampling*, which is a type of nonprobability sampling.

Cluster sampling can be combined with stratification to reduce the variance of estimators. Stratification is the process of dividing the units, or clusters, in a population into strata, or groups. Although similar to cluster sampling, in which several clusters are selected to be in the sample, stratified sampling entails selecting independent samples within every stratum. Stratified sampling typically reduces variance of estimators by forcing the sample to include representatives of all strata. In a survey of high school students in Iowa, the school districts could be stratified into the twelve Area Education Agencies within the state, and then a multistage sample of students could be selected within each Area Education Agency. Stratified cluster samples aim to combine the convenience of cluster sampling with precise estimation produced by stratification.

—Michael D. Larsen

See also Probability Sampling

Further Reading

- Cochran, W.G. (1977). *Sampling techniques*. New York: Wiley.
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.
- Scheaffer, R. L., Mendenhall, W., III, & Ott, L. R. (1995). *Elementary survey sampling* (6th ed.). Belmont, CA: Duxbury Press.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261–304.

This article demonstrates the use of stratified and **cluster sampling** to draw a sample from the U.S. Census Archives for California in 1880. This is a useful exercise for courses in research design and also provides an efficient method for taking samples for historical research. With relatively little effort spent on data collection and data entry, useful knowledge about California in 1880 was acquired pertaining to marriage patterns, migration patterns, occupational status, and categories of race and ethnicity.

COCHRAN Q TEST

In a randomized complete block design, it is often of interest to examine a set of c binary responses pertaining to the levels of some treatment condition provided either by a sample of participants used as their own controls to make these assessments or by a sample of c matched participants randomly assigned to each treatment level as members of a block. In 1950, the statistician William G. Cochran developed a test for the differences in the proportions of “success” among the treatment levels in c related groups in which the repeated measurements provided by each participant under c different conditions or the measurements provided by the c homogeneous participants in each block are binary responses, with success coded 1 and “failure” coded 0.

The Cochran Q test is a dichotomous data counterpart of another nonparametric procedure, the Friedman rank test, in which the responses to the c treatment levels within each block are ranked. Both these procedures are competitors of the classic two-way ANOVA randomized block F test, in which the responses to each treatment are measured on an interval or ratio scale and the assumption of underlying normality of the different treatment groups is made.

Motivation

Cochran's Q test has enjoyed widespread use in behavioral, business, educational, medical, and social science research when it may be desirable to evaluate possible significance of differences in the proportion of successes over several treatment groups.

As an example of a study in which each participant offers repeated measurements, one each for the c levels of the treatment or condition under evaluation, the efficacy of c different drugs prescribed for providing relief from some chronic condition may be investigated. Other examples of this type could involve taste testing wines or other consumer preference studies in which each of c items is classified as "acceptable" or "not acceptable." As an example of a randomized complete block experiment in which the c homogeneous members of a block are randomly assigned, one each, to the c levels of the treatment, an educator may wish to form sets of homogeneous blocks of students and randomly assign the members of the block to each of c learning methods with the goal of assessing whether there are significant differences among the learning methods based on the proportions of successes observed.

Development

The layout for the dichotomous responses from a sample of either r participants or r blocks of matched participants over c levels of a treatment condition is shown in Table 1.

Cochran's Q test statistic is given by

$$Q = \frac{(c-1)(c \sum_{j=1}^c x_{.j}^2 - N^2)}{cN - \sum_{i=1}^r x_i^2},$$

Table 1 Data Layout for the Cochran Q Test

Block	Treatments				Totals
	1	2	...	c	
1	x_{11}	x_{12}	...	x_{1c}	$x_{1.}$
2	x_{21}	x_{22}	...	x_{2c}	$x_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	x_{r1}	x_{r2}	...	x_{rc}	$x_{r.}$
Totals	$x_{.1}$	$x_{.2}$	...	$x_{.c}$	N
Proportion of "Success"	$\hat{p}_{.1} = \frac{x_{.1}}{r}$	$\hat{p}_{.2} = \frac{x_{.2}}{r}$	...	$\hat{p}_{.c} = \frac{x_{.c}}{r}$	

Notes: c = the number of treatment groups (i.e., columns); r = the number of blocks (i.e., rows of subjects); x_{ij} = the binary response ("success" = 1, "failure" = 0) for the j th treatment in the i th block; $x_{.j}$ = the total number of successes for treatment j ; $x_{i.}$ the total number of successes for block i ;

$N = \sum_{j=1}^c x_{.j} = \sum_{i=1}^r x_{i.}$ the total number of successes.

where

c is the number of treatment groups (i.e., columns),

r is the number of blocks (i.e., rows of subjects),

x_{ij} is the binary response (success = 1, failure = 0) for the j th treatment in the i th block,

$x_{.j}$ is the total number of successes for treatment j ,

$x_{i.}$ is the total number of successes for block i , and

$N = \sum_{j=1}^c x_{.j} = \sum_{i=1}^r x_{i.}$ = the total number of successes.

As the number of "discriminating" blocks (i.e., those in which outcomes for the c treatments are not either all 1s or all 0s) gets large, the Cochran Q test statistic is approximated by a chi-square distribution with $c - 1$ degrees of freedom. The decision rule is to reject the null hypothesis of no differences (i.e., no treatment effect), $H_0: p_{.1} = p_{.2} = \dots = p_{.c}$, if at the α level of significance, $Q > \chi^2_{\alpha, (c-1)}$.

The researchers M. W. Tate and S. W. Brown recommended that there be at least seven such discriminating blocks in studies for which four treatment levels are involved in order to use the chi-square distribution as an approximation to Q , and both they and the statistician K. D. Patil provided tables of critical

values for an exact test of Q that are particularly useful when the numbers of discriminating blocks and treatment levels are small.

A Posteriori Comparisons

If the null hypothesis is rejected, the researchers L. A. Marascuilo and M. McSweeney suggested a multiple comparison procedure that permits a post hoc evaluation of all pairwise differences among the c treatment groups. With an experimentwise error rate α , each of the possible $c(c-1)/2$ pairwise comparisons is made, and the decision rule is to declare treatment j different from treatment j' if

$$\left| \frac{x_{.j}}{r} - \frac{x_{.j'}}{r} \right| > \sqrt{\chi_{\alpha, (c-1)}^2} \cdot \sqrt{\frac{2(cN - \sum_{i=1}^r x_i^2)}{r^2 c(c-1)}}.$$

That is, treatment j and treatment j' are declared significantly different if $|\hat{p}_{.j} - \hat{p}_{.j'}|$, the absolute difference in the sample proportions of success, exceeds a critical range given by the product of

$$\sqrt{\chi_{\alpha, (c-1)}^2} \text{ and } \sqrt{\frac{2(cN - \sum_{i=1}^r x_i^2)}{r^2 c(c-1)}}.$$

Applying Cochran's Q Test

Consider the following hypothetical example: Suppose a group of 16 faculty members is asked to evaluate the resumes of each of four candidates who have interviewed at the college for the position of dean. Each of the 16 faculty is to assign scores of 1 (would recommend) or 0 (would not recommend) to each of the four candidates. The results are displayed in Table 2. Note that one of the professors (*JL*) did not recommend any of the four candidates as

qualified for the position of dean while one of the professors (*AO*) recommended all the candidates. The other 14 faculty were able to discriminate among the applicants and provide favorable recommendations to one or more of these candidates.

For these data, the Cochran Q test is used to test the null hypothesis of no treatment effect (that is, each of the four candidates is preferred equally, and any differences in the observed proportions are due to chance),

$$H_0: p_{.1} = p_{.2} = \cdots = p_{.c},$$

against the general alternative that a treatment effect is present (that is, there are real differences in the proportions; at least one of the candidates is preferred differently from the others):

$$H_1: \text{Not all } p_j \text{ are equal (where } j = 1, 2, \dots, c).$$

From Table 2, it is observed that $c = 4$, $r = 16$, $N = 31$, and the x_j and x_i are summarized in the column

Table 2 Hypothetical Results of Evaluating Job Candidates

Raters	Candidates				Totals
	1 = IC	2 = PE	3 = NP	4 = PS	
1 = HB	0	0	1	0	1
2 = RB	0	1	1	1	3
3 = KC	1	0	1	1	3
4 = MD	0	0	1	1	2
5 = EF	0	0	1	1	2
6 = KH	0	0	1	0	1
7 = BK	0	0	0	1	1
8 = JK	0	0	1	1	2
9 = JL	0	0	0	0	0
10 = RM	0	0	0	1	1
11 = AO	1	1	1	1	4
12 = RP	0	0	1	1	2
13 = JT	0	0	1	1	2
14 = JW	1	0	1	1	3
15 = JY	0	0	1	1	2
16 = RZ	1	0	1	0	2
Totals	4	2	13	12	31
Proportions	$\hat{p}_{.1} = 0.2500$	$\hat{p}_{.2} = 0.1250$	$\hat{p}_{.3} = 0.8125$	$\hat{p}_{.4} = 0.7500$	

and row totals. Cochran's Q test statistic is then computed as follows:

$$\begin{aligned}
 Q &= \frac{(c-1) \left(c \sum_{j=1}^c x_{.j}^2 - N^2 \right)}{cN - \sum_{i=1}^r x_{i.}^2} \\
 &= \frac{(3)[(4)(4^2 + 2^2 + 13^2 + 12^2) - 31^2]}{(4)(31) - (1^2 + 3^2 + 3^2 + \dots + 2^2)} \\
 &= \frac{1113}{49} = 22.714.
 \end{aligned}$$

Since the Cochran Q test statistic is approximated by a chi-square distribution with $c - 1$ degrees of freedom, using an $\alpha = 0.05$ level of significance, the decision rule is to reject the null hypothesis of no differences ($H_0 : p_{.1} = p_{.2} = \dots = p_{.c}$) if $Q > \chi_{0.05, (c-1=3)}^2 = 7.815$, the upper-tailed critical value under the chi-square distribution with 3 degrees of freedom. The null hypothesis is rejected, and it is concluded that significant differences in preferences for the candidates exist.

Given that the null hypothesis is rejected, to determine which candidate(s) significantly stand out, post hoc evaluations of all $(4)(3)/2 = 6$ pairwise differences among the 4 candidates are made. The critical range for these pairwise comparisons is

$$\begin{aligned}
 &\sqrt{\chi_{\alpha, (c-1)}^2} \cdot \sqrt{\frac{2(cN - \sum_{i=1}^r x_{i.}^2)}{r^2 c (c-1)}} = \sqrt{7.815} \\
 &\cdot \sqrt{\frac{(2)[(4)(31) - (1^2 + 3^2 + 3^2 + \dots + 2^2)]}{(16^2)(4)(3)}} \\
 &= 0.499.
 \end{aligned}$$

From the sample proportions of success summarized at the bottom of Table 2, the pairwise comparisons are evaluated in Table 3.

From Table 3 it is clear that candidates 3 and 4 are each significantly preferred to candidates 1 and 2. Nevertheless, the difference in preference between

Table 3 Post Hoc Pairwise Comparisons of Four Candidates

Candidates	$ \hat{p}_{.j} - \hat{p}_{.j'} $	Critical Range	Decision Rule
1 vs. 2	$ \hat{p}_{.1} - \hat{p}_{.2} = 0.1250$	0.499	Not significant
1 vs. 3	$ \hat{p}_{.1} - \hat{p}_{.3} = 0.5625$	0.499	Significant
1 vs. 4	$ \hat{p}_{.1} - \hat{p}_{.4} = 0.5000$	0.499	Significant
2 vs. 3	$ \hat{p}_{.2} - \hat{p}_{.3} = 0.6875$	0.499	Significant
2 vs. 4	$ \hat{p}_{.2} - \hat{p}_{.4} = 0.6250$	0.499	Significant
3 vs. 4	$ \hat{p}_{.3} - \hat{p}_{.4} = 0.0625$	0.499	Not significant

candidates 1 and 2 is not significant, and more important here, the difference in preference between candidates 3 and 4 is not significant. The recommendation from the faculty regarding the appointment of the dean would be for candidate 3 or 4. Other criteria would be needed to finalize the decision process.

Discussion

The Cochran Q test can also be viewed as a c sample generalization of McNemar's test for significance of change in two proportions based on related samples.

It is essential to a good data analysis that the appropriate statistical procedure be applied to a specific situation. When comparing differences in c proportions based on related samples, where the responses in each of the blocks are simply binary rather than ranked or measured on some interval or ratio scale, Cochran's Q test should be selected.

Statisticians P. P. Ramsey and P. H. Ramsey investigated the minimum block sizes or sample sizes needed to apply Cochran's Q test, and biostatisticians S. Wallenstein and A. Berger studied the power properties of the test.

Conclusions

The Cochran Q test is quick and easy to perform. The only assumptions are that the outcomes for each

response are binary and that either the participants providing these c binary responses are randomly selected or the blocks of homogeneous participants examining the c treatment levels are randomly selected.

When evaluating the worth of a statistical procedure, statistician John Tukey defined *practical power* as the product of statistical power and the utility of the statistical technique. Based on this, the Cochran Q test enjoys a high level of practical power under many useful circumstances.

—Mark L. Berenson

See also Repeated Measures Analysis of Variance

Further Reading

- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256–266.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole.
- Patil, K. D. (1975). Cochran's Q test: Exact distribution. *Journal of the American Statistical Association*, 70, 186–189.
- Ramsey, P. P., & Ramsey, P. H. (1981). Minimum sample sizes for Cochran's test. Retrieved May 9, 2006, from http://www.amstat.org/sections/srms/Proceedings/papers/1981_146.pdf
- Tate, M. W., & Brown, S. M. (1970). Note on Cochran's Q test. *Journal of the American Statistical Association*, 65, 155–160.
- Tukey, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics*, 1, 31–48.
- Wallenstein, S., & Berger, A. (1981). On the asymptotic power of tests for comparing k correlated proportions. *Journal of the American Statistical Association*, 76, 114–118.

COEFFICIENT ALPHA

Two criteria are usually used to assess the quality of a test: reliability and validity. There are several different types of reliability, including test-retest, interrater, alternate form, and internal consistency. These tell us how much of the variance in the test scores is due to true differences between people and how much is due

to time, rater, form, and items, respectively. The first three types of reliability can be assessed by calculating a simple correlation. Test-retest reliability can be assessed by correlating total scores from two different testing times. Interrater reliability can be assessed by correlating the scores given by two different raters. Alternate form reliability can be assessed by correlating total scores from different forms of the test. Each of these correlations tells us the reliability of a single measurement: when we use one test, one rater, or one form. However, when it comes to internal consistency, we cannot simply use a correlation, because we very rarely use just a single item. To assess internal consistency, we therefore need to use a formula that tells us the reliability of the sum of several different measurements. The formula we use most often is coefficient alpha.

Other Names for Coefficient Alpha

Coefficient alpha, developed by Lee J. Cronbach in 1951, is mathematically identical to several other formulas developed to assess the reliability of total scores. These include Cyril J. Hoyt's estimate of reliability from 1941 and, if items are dichotomous, G. Frederik Kuder and M. W. Richardson's Formula 20. If all items on a test have the same variance, coefficient alpha is also equal to Kuder and Richardson's Formula 21 and the Spearman-Brown prophecy formula. Coefficient alpha is the most general of the formulas, which probably explains why it is the most commonly used version of the formula. Not surprisingly, it is also known as Cronbach's alpha.

Uses

Coefficient alpha is usually used to assess the reliability of total test scores when a test is made up of many items. It is therefore usually thought of as a measure of internal consistency. However, coefficient alpha can also be used to assess the reliability of other types of total scores. For example, if three letters of reference are solicited when evaluating applicants, coefficient alpha can be used to assess the reliability of total scores from those three letters. Alternatively, in a

diary study, respondents may answer the same question every day for a month. The reliability of the total of those scores can be estimated with coefficient alpha. This formula works equally well whether the researcher calculates the sum of the measurements or the average of the measurements: The formula will result in the exact same number.

Formula

The population value of coefficient alpha is calculated as follows:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_{Total}^2} \right),$$

where

k is the number of measurements,

σ_i^2 is the population variance of the i th measurement, and

σ_{Total}^2 is the population variance of total scores on the k measurements.

However, we rarely have population data. Instead, we estimate the population value of coefficient alpha from the sample data of people we actually measured in our study. The sample value of coefficient alpha is calculated by substituting sample variances for population variances, as follows:

$$\hat{\alpha} = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_{Total}^2} \right),$$

where

k is the number of measurements,

s_i^2 is the sample variance of the i th measurement, and

s_{Total}^2 is the sample variance of total scores on the k measurements.

Example of the Hand Calculation

Consider the following example, in which four students completed three items on a test. Scores for each item appear in the columns marked *Item*.

<i>Student</i>	<i>Item 1</i>	<i>Item 2</i>	<i>Item 3</i>	<i>Total</i>
1	3	4	1	8
2	2	2	1	5
3	9	8	4	21
4	6	2	0	8

To calculate coefficient alpha by hand, you must first calculate the variances of each measurement and of the total scores. These are as follows.

	<i>Item 1</i>	<i>Item 2</i>	<i>Item 3</i>	<i>Total</i>
Sample variance	10	8	3	51

Now we can calculate coefficient alpha:

$$\begin{aligned} \hat{\alpha} &= \frac{3}{3-1} \left(1 - \frac{10+8+3}{51} \right) \\ &= \frac{3}{2} \left(1 - \frac{21}{51} \right) = .88. \end{aligned}$$

Hand calculation of coefficient alpha can be quite tedious, especially if there are a large number of respondents and a large number of items. Fortunately, statistical packages like SPSS make these calculations easy.

Example of the Calculation Using SPSS

In SPSS, calculation of coefficient alpha is easy. Enter your data into the Data Editor, with one respondent per row. Then click on the Analyze menu, select Scale from the drop-down menu, and select Reliability Analysis from the side menu. This will bring you to the Reliability Analysis dialogue box. Select your measurements and move them across to the Items box. Click OK. From the same data as in the hand calculation above, Figure 1 was produced.

Reliability**Case Processing Summary**

		N	%
Cases	Valid	4	100.0
	Exclude ^a	0	.0
	Total	4	100.0

a. Listwise deletion based on all variables in the procedure

Reliability Statistics

Cronbach's Alpha	N of items
.882	3

Figure 1 SPSS Output for Coefficient Alpha

The first table in Figure 1 shows that there were four respondents. The second table shows that coefficient alpha is .882 and that there were three items on the test.

Interpretation

Like other reliability coefficients, coefficient alpha tells you the proportion of total score variance that is due to true differences among respondents. It also tells you the expected correlation between total scores on your k measurements and total scores on a hypothetical set of k other measurements that were designed to measure the same construct. In 1963, Lee Cronbach, Nageswari Rajaratnam, and Goldine C. Glesser showed that coefficient alpha provides an unbiased estimate of the true reliability of a test if the test user assumes that the items were randomly sampled from a set of other items that could have been used to measure the same construct.

The most common misinterpretation of coefficient alpha is to assume that the level of internal consistency tells the test user something about the other types of reliability. Four types of reliability have been mentioned already: test-retest reliability, interrater reliability, alternate form reliability, and internal consistency reliability. It is possible for a test to have high

internal consistency but low test-retest reliability. Consider, for example, a test of mood. It is also possible for a test to have high test-retest reliability but low internal consistency. Consider, for example, a test that consists of three items: hat size, the last four digits of your phone number, and number of children. Over a two-week period, total scores on this test are likely to be very stable, but these items will have very low internal consistency. Internal consistency is easy to estimate, but it does not substitute for estimates of other types of reliability. Research on other types of reliability is still needed.

Inferential Procedures for Coefficient Alpha

Several inferential procedures for coefficient alpha have been developed. These include confidence intervals, comparisons between independent alpha coefficients, and comparisons between dependent alpha coefficients. Leonard S. Feldt, David J. Woodruff, and Fathi A. Salih wrote a summary of these procedures in 1987. Unfortunately, these inferential tests have fairly restrictive assumptions—that all measurements have equal variances and equal covariances—and Kimberly A. Barchard and A. Ralph Hakstian in two papers in 1997 showed that these procedures are not robust to violation of these assumptions. Hakstian and Barchard's initial attempts to develop a correction for violation of this assumption were only partially successful, and therefore the results of these inferential procedures should be considered tentative.

—Kimberly A. Barchard

See also Classical Test Theory; Reliability Theory

Further Reading

- Barchard, K. A., & Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational & Psychological Measurement*, 57, 893–905.
- Barchard, K. A., & Hakstian, A. R. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Research*, 32, 169–191.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.

- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11, 93–103.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 20, 1–22.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Lopez, M. N., Lazar, M. D., & Oh, S. (2003). Psychometric properties of the Hooper Visual Organization Test. *Assessment*, 10(1), 66–70.

Coefficient alpha is one of many measures to assess the internal consistency or reliability of a test. In this study, the authors present internal consistency and interrater reliability coefficients and an item analysis using data from a sample ($N = 281$) of “cognitively impaired” and “cognitively intact” patients and patients with undetermined cognitive status. Coefficient alpha for the Visual Organization Test (VOT) total sample was .882. Of the 30 items, 26 were good at discriminating among patients. Also, the interrater reliabilities for three raters (.992), two raters (.988), and one rater (.977) were excellent. The authors’ conclusion is that the judgmental scoring of the VOT does not interfere significantly with its clinical utility and that the VOT is a psychometrically sound test.

COEFFICIENTS OF CORRELATION, ALIENATION, AND DETERMINATION

The coefficient of correlation evaluates the similarity of two sets of measurements (i.e., two dependent variables) obtained on the same observations. The coefficient of correlation indicates the amount of information common to two variables. This coefficient takes values between -1 and $+1$ (inclusive).

A value of 0 indicates that the two series of measurement have nothing in common. A value of $+1$ says that the two series of measurements are measuring the same thing. A value of -1 says that the two measurements are measuring the same thing, but one measurement varies inversely with the other.

The squared correlation gives the proportion of common variance between two variables and is also called the *coefficient of determination*. Subtracting the coefficient of determination from the unity gives the proportion of variance not shared between two variables, a quantity also called the *coefficient of alienation*.

The coefficient of correlation measures only the *linear* relationship between two variables, and its value is very sensitive to outliers. Its significance can be tested with an F or a t test. The coefficient of correlation always overestimates the intensity of the correlation in the population and needs to be “corrected” in order to provide a better estimation. The corrected value is called “shrunk” or “adjusted.”

Notations and Definition

We have S observations, and for each observation, we have two measurements, denoted W and Y , with respective means M_W and M_Y . For each observation, we define the cross product as the product of the deviations of each variable to its mean. The sum of these cross products, denoted SCP_{WY} , is computed as

$$SCP_{WY} = \sum_s (W_s - M_W)(Y_s - M_Y). \quad (1)$$

The sum of the cross products reflects the association between the variables. When the deviations tend to have the same sign, they indicate a positive relationship, and when they tend to have different signs, they indicate a negative relationship. The average value of the SCP_{WY} is called the covariance (cov), and just like the variance, the covariance can be computed by dividing by S or $(S - 1)$:

$$cov_{WY} = \frac{SCP}{\text{Number of Observations}} = \frac{SCP}{S}. \quad (2)$$

The covariance reflects the association between the variables, but it is expressed in the original units of measurement. In order to eliminate them, the covariance is normalized by division by the standard deviation of each variable (σ). This defines the coefficient of correlation, denoted $r_{W.Y}$, which is equal to

$$r_{W.Y} = \frac{COV_{WY}}{\sigma_W \sigma_Y}. \quad (3)$$

Rewriting the previous formula gives a more practical formula:

$$r_{W.Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}. \quad (4)$$

An Example: Correlation Computation

We illustrate the computation for the coefficient of correlation with the following data, describing the values of W and Y for $S = 6$ subjects:

$$\begin{aligned} W_1 = 1 \quad W_2 = 3 \quad W_3 = 4 \quad W_4 = 4 \quad W_5 = 5 \quad W_6 = 7, \\ Y_1 = 16 \quad Y_2 = 10 \quad Y_3 = 12 \quad Y_4 = 4 \quad Y_5 = 8 \quad Y_6 = 10. \end{aligned}$$

Step 1: Computing the sum of the cross products

First compute the means of W and Y :

$$\begin{aligned} M_W &= \frac{1}{S} \sum_{s=1}^S W_s = \frac{24}{6} = 4 \\ &\text{and} \\ M_Y &= \frac{1}{S} \sum_{s=1}^S Y_s = \frac{60}{6} = 10. \end{aligned}$$

The sum of the cross products is then equal to

$$\begin{aligned} SCP_{YW} &= \sum_s (Y_s - M_Y)(W_s - M_W) \\ &= (16 - 10)(1 - 4) + (10 - 10)(3 - 4) \\ &\quad + (12 - 10)(4 - 4) + (4 - 10)(4 - 4) + (8 - 10)(5 - 4) \\ &\quad + (10 - 10)(7 - 4) \\ &= (6 \times -3) + (0 \times -1) + (2 \times 0) + (-6 \times 0) \\ &\quad + (-2 \times 1) + (0 \times 3) \\ &= -18 + 0 + 0 + 0 - 2 + 0 \\ &= -20. \end{aligned} \quad (5)$$

The sum of squares of W_s is obtained as

$$\begin{aligned} SS_W &= \sum_{s=1}^S (W_s - M_W)^2 \\ &= (1 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (4 - 4)^2 \\ &\quad + (5 - 4)^2 + (7 - 4)^2 \\ &= (-3)^2 + (-1)^2 + 0^2 + 0^2 + 1^2 + 3^2 \\ &= 9 + 1 + 0 + 0 + 1 + 9 \\ &= 20. \end{aligned} \quad (6)$$

The sum of squares of Y_s is

$$\begin{aligned} SS_Y &= \sum_{s=1}^S (Y_s - M_Y)^2 \\ &= (16 - 10)^2 + (10 - 10)^2 + (12 - 10)^2 \\ &\quad + (4 - 10)^2 + (8 - 10)^2 + (10 - 10)^2 \\ &= 6^2 + 0^2 + 2^2 + (-6)^2 + (-2)^2 + 0^2 \\ &= 36 + 0 + 4 + 36 + 4 + 0 \\ &= 80. \end{aligned} \quad (7)$$

Step 2: Computing r_{WY}

The coefficient of correlation between W and Y is equal to

$$\begin{aligned} r_{W.Y} &= \frac{\sum_s (Y_s - M_Y)(W_s - M_W)}{\sqrt{SS_Y \times SS_W}} \\ &= \frac{-20}{\sqrt{80 \times 20}} = \frac{-20}{\sqrt{1600}} = \frac{-20}{40} \\ &= -.5. \end{aligned} \quad (8)$$

We can interpret this value of $r = -.5$ as an indication of a negative linear relationship between W and Y .

Some Properties of the Coefficient of Correlation

The coefficient of correlation is a number without a unit. This occurs because of dividing the units of the numerator by the same units in the denominator. Hence, the coefficient of correlation can be used to compare outcomes across different variables. The

magnitude of the coefficient of correlation is always smaller than or equal to 1. This happens because the numerator of the coefficient of correlation (see Equation 4) is always smaller than or equal to its denominator (this property follows from the Cauchy-Schwartz inequity). A coefficient of correlation equal to +1 or -1 indicates that a plot of the observations will show that they are positioned on a line.

The squared coefficient of correlation gives the *proportion of common variance* between two variables, also called the coefficient of determination. In our example, the coefficient of determination is equal to $r^2_{WY} = .25$. The proportion of variance not shared between the variables, or the coefficient of alienation, is, for our example, equal to $1 - r^2_{WY} = .75$.

Interpreting Correlation

Linear and Nonlinear Relationship

The coefficient of correlation measures only linear relationships between two variables and will miss nonlinear relationships. For example, Figure 1 displays a perfect nonlinear relationship between two variables (i.e., the data show a U-shaped relationship,

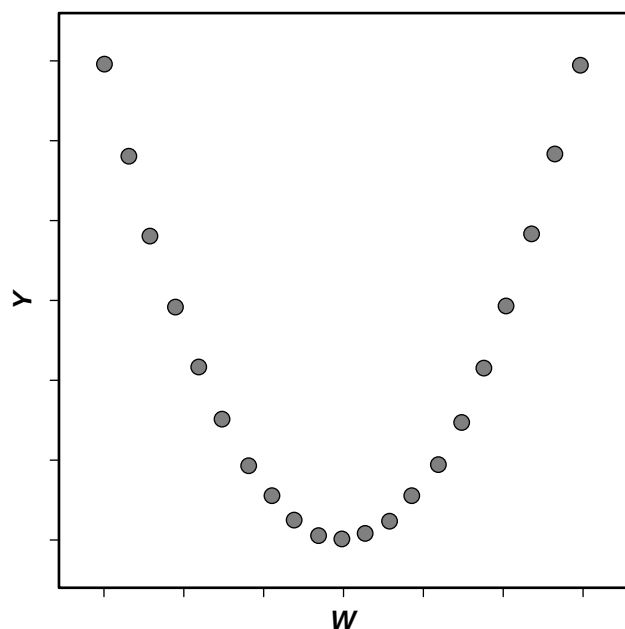


Figure 1 A Perfect Nonlinear Relationship With Coefficient of Correlation = 0

with Y being proportional to the square of W), but the coefficient of correlation is equal to 0.

The Effect of Outliers

Observations far from the center of the distribution contribute a lot to the sum of the cross products. At the extreme, in fact, as illustrated in Figure 2, one extremely deviant observation (often called an *outlier*) can dramatically influence the value of r .

Geometric Interpretation:

The Coefficient of Correlation Is a Cosine

Each set of observations can also be seen as a vector in an S dimensional space (one dimension per observation). Within this framework, the correlation is equal to the cosine of the angle between the two vectors after they have been centered by subtracting their respective means. For example, a coefficient of correlation of $r = -.50$ corresponds to a 150-degree angle.

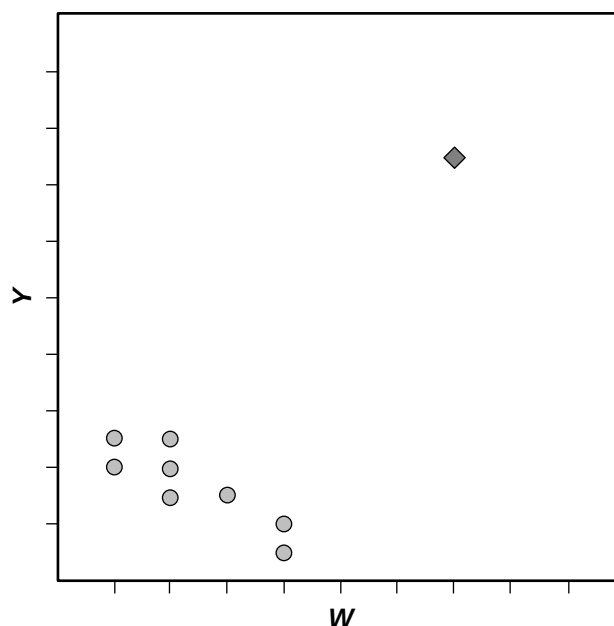


Figure 2 The Dangerous Effect of Outliers on the Value of the Coefficient of Correlation

Note: The correlation of the set of points represented by the circles is equal to $-.87$, but when the point represented by the diamond is added to the set, the correlation is equal to $+.61$. This shows that an outlier can completely determine the value of the coefficient of correlation.

A coefficient of correlation of 0 corresponds to a right angle, and therefore two uncorrelated variables are called *orthogonal* (which is derived from the Greek word for a right angle).

Correlation and Causation

The fact that two variables are correlated does not mean that one variable causes the other one: Correlation is not causation. For example, in France, the number of Catholic churches, as well as the number of schools, in a city is highly correlated with the number of people with cirrhosis of the liver, the number of teenage pregnancies, and the number of violent deaths. Does that mean that churches and schools are sources of vice and that newborns are murderers? Here, in fact, the observed correlation is due to a third variable, namely, the size of the cities: The larger a city, the larger the number of churches, schools, alcoholics, and so on. In this example, the correlation between number of churches or schools and alcoholics is called a spurious correlation because it reflects only their mutual correlation with a third variable (i.e., size of the city).

Testing the Significance of r

A null hypothesis test for r can be performed using an F statistic obtained as follows:

$$F = \frac{r^2}{1 - r^2} \times (S - 2). \quad (9)$$

When the null hypothesis is true (and when the normality assumption holds), this statistic is distributed as a Fisher's F with $v_1 = 1$ and $v_2 = S - 2$ degrees of freedom. An equivalent test can be performed using $t = \sqrt{F}$, which is distributed under H_0 as a Student's distribution with $v = S - 2$ degrees of freedom.

For our example, we find that

$$\begin{aligned} F &= \frac{.25}{1 - .25} \times (6 - 2) = \frac{.25}{.75} \times 4 \\ &= \frac{1}{3} \times 4 = \frac{4}{3} = 1.33. \end{aligned}$$

The probability of finding such a value under H_0 is found using an F distribution with $v_1 = 1$ and $v_2 = 3$ and is equal to $p \approx .31$. Such a value does not lead to rejecting H_0 .

Estimating the Population Correlation: Shrunk and Adjusted r

The coefficient of correlation is a *descriptive* statistic that always overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. In order to obtain a better estimate of the population, the value r needs to be corrected. As suggested earlier, the corrected value of r goes under various names: corrected r , shrunk r , or adjusted r (there are some subtle differences between these terms, but we will ignore them here), and we denote it by $\tilde{r}^2$. Several correction formulas are available; the one used most often estimates the value of the population correlation as

$$\tilde{r}^2 = 1 - \left[(1 - r^2) \left(\frac{S - 1}{S - 2} \right) \right]. \quad (10)$$

For our example, this gives

$$\begin{aligned} \tilde{r}^2 &= 1 - \left[(1 - r^2) \left(\frac{S - 1}{S - 2} \right) \right] \\ &= 1 - \left[(1 - .25) \times \frac{5}{4} \right] \\ &= 1 - \left[.75 \times \frac{5}{4} \right] = 0.06. \end{aligned}$$

With this formula, we find that the estimation of the population correlation drops from $r = -.50$ to $\tilde{r} = -\sqrt{\tilde{r}^2} = -\sqrt{.06} = -.24$.

—Hervé Abdi

See also Correlation Coefficient; Multiple Correlation Coefficient; Spurious Correlation

Further Reading

- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the social sciences*. Hillsdale, NJ: Erlbaum.
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. New York: Harcourt-Brace.

COGNITIVE ABILITIES TEST

The Cognitive Abilities Test (CogAT), published by Riverside Publishing (www.riverpub.com), is a group-administered test appraising developed reasoning abilities. Its 11 test levels span Kindergarten through Grade 12. CogAT is the contemporary successor to the Lorge-Thorndike Intelligence Tests. In the spring of 2000, the sixth edition of the test was conormed with the Iowa Tests of Basic Skills (Grades K–8) and the Iowa Tests of Educational Development (Grades 9–12). The standardization sample consisted of more than 180,000 students in public, Catholic, and private non-Catholic schools. When administered with either of the Iowa tests, CogAT discrepancies between observed and predicted achievement scores may be obtained for each examinee.

CogAT measures abstract reasoning abilities in the three major symbol systems used to communicate knowledge in schools: verbal, quantitative, and figural/spatial. It reports both age- and grade-normed scores for all three reasoning abilities, plus a composite score. Verbal, quantitative, and nonverbal reasoning scores are estimated by two subtests in the Primary Edition (Grades K–2) and by three subtests in the Multilevel Edition (Grades 3–12). Items in the Primary Edition are paced by the teacher and require no reading. Tests in the Multilevel Edition require some reading and are administered with time limits. Testing time for the Multilevel Edition is 90 minutes.

The test authors suggest that the most important uses of CogAT scores are (a) to guide efforts to adapt instruction to the needs and abilities of students, (b) to provide a measure of cognitive development that usefully supplements achievement test scores and teacher

grades, and (c) to identify for further study those students whose predicted levels of achievement differ markedly from their observed levels of achievement. The first use is supported through several teacher guides and a Web-based system for matching the level and pattern of a student's CogAT scores to specific instructional recommendations (see www.cogat.com). Recommendations are based on recent summaries of the aptitude-by-treatment interaction literature. That literature shows that reasoning abilities in the symbol systems used to communicate new knowledge are among the most important aptitudes for success in school and thus interact with variations in instructional methods.

CogAT manuals provide considerable assistance in avoiding common mistakes when interpreting test scores. In addition to the *Research Handbook* (104 pages) and *Norms Booklet* (128 pages), there are an extensive *Interpretive Guide for Teachers and Counselors* (166 pages) and an *Interpretive Guide for School Administrators* (134 pages). A *Short Guide for Teachers* is available at no charge on the CogAT Web site (www.cogat.com). Scores on Form 6 are flagged if they appear unsound in any of nine different ways. One of the innovative features of CogAT6 is the introduction of confidence intervals for each score. The confidence intervals are based both on the conditional standard error of measurement and an estimate of fit. In this way, users are warned if the response pattern on a battery is aberrant for a particular examinee.

—David F. Lohman

See also Cognitive Psychometric Assessment; Kingston Standardized Cognitive Assessment

Further Reading

- Corno, L., Cronbach, L. J., Lohman, D. F., Kupermintz, H., Mandinach, E. B., Porteus, A., et al. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Hillsdale, NJ: Erlbaum.
- Lohman, D. F. (2005). The role of non-verbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly*, 49, 111–138.
- Lohman, D. F. (in press). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*.

Lorge, I., & Thorndike, R. L. (1954). *The Lorge-Thorndike Intelligence Tests*. Boston: Houghton Mifflin.

CogAT recent research links: www.cogat.com, <http://faculty.education.uiowa.edu/dlohman>

COGNITIVE PSYCHOMETRIC ASSESSMENT

Cognitive psychometric assessment (CPA) is the assessment of psychological traits such as abilities, interests, and dispositions based on data from instruments such as questionnaires and tests that are analyzed with extensions of psychometric models. The models used for CPA are *cognitive psychometric models* (CPM) that contain discrete or continuous latent variables; their theoretical foundations are in the areas of item response theory and structural equation modeling.

Specifically, CPMs include parameters that operationalize components of mental processes or faculties whose existence can be justified through theories that draw on cognitive psychology. The parameters are provided by specialists and are typically collected in so-called Q-matrices, whose entries may be binary (i.e., indicating the absence or presence of a component) or ordinal (i.e., indicating the degree to which a component is present). In many models, such as the *Linear Logistic Test Model*, the *DINA* and *NIDA Models*, or the *Rule-Space Methodology*, these parameters are fixed whereas in some models, such as the *Reparametrized Unified Model* or *Fusion Model*, they are subject to empirical updating.

Alternatively, the structure of CPMs can reflect a specific combination of components. In this case, the number of components is provided by the specialists, and the model structure is chosen to match the way in which examinees engage in mental processes to respond to items. For example, in the *Multidimensional Rasch Model for Learning and Change*, deficiencies in one component can be compensated for by strengths in another component, which is also known as a compensatory or disjunctive model. In contrast, in the *Multidimensional Logistic Trait Model*, deficiencies in one component cannot be compensated for by strengths

in another component; therefore, such a model is known as a noncompensatory or conjunctive model.

In order to conduct CPA successfully in practice, a well-developed theory about the cognitive processes underlying item responses is necessary. Experience has shown that this is much easier to accomplish for tasks that can be easily decomposed into constituent elements such as mathematical addition and subtraction but is much harder for complex reasoning and problem-solving tasks such as reading comprehension. Such a theory entails a detailed description of how the tasks that are utilized provide the kinds of empirical evidence that are needed to make the kinds of inferences that are desired. Moreover, the successful application of CPMs in practice requires that sufficiently large sample sizes be available for model calibration to achieve convergence for parameter estimation routines and to achieve reliable classifications. The process of understanding how response patterns influence the estimation of CPMs is just beginning, however, and more empirical investigation to develop practical recommendations for their use is needed.

—André A. Rupp

See also Cognitive Abilities Test; Kingston Standardized Cognitive Assessment

Further Reading

- de la Torre, J., & Douglas, J. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333–353.
- Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32, 277–294.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Hartz, S. M. (2002). *A Bayesian guide for the Unified Model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.
- Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement*, 25, 258–272.

- Maris, E. (1995). Psychometric latent response models. *Psychometrika*, 60, 523–547.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and perspectives*, 1, 3–62.
- Rupp, A. A. (2006). *The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. Manuscript submitted for publication.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Erlbaum.

COHEN'S KAPPA

Cohen's kappa statistic was developed to correct for the problem of inflated percent agreement statistics that occur when marginal values on the variables being compared are unevenly distributed. Kappa is typically used with nominal level variables and is often seen in situations in which two independent raters have the task of classifying an object as belonging to a single level of a nominal variable.

For example, consider the following data set in which two raters, Steve and Damian, are asked to read 100 school mission statements. For each mission statement, they are to make a judgment as to the dominant purpose of schooling set forth in the mission statement. Each school may have only one dominant theme, and the theme should fit into one of the following categories: (a) social, (b) cognitive, (c) civic, or (d) emotional. The results of the rater classifications are shown in Table 1, where the values are reported in terms of percentages (e.g., the value of .05 in the social box indicates that 5 out of 100 schools were classified as having a social purpose as their dominant theme).

The marginal totals indicate the percentage of ratings assigned to each category for each rater. In this example, Steve classified 61% of the 100 mission statements as belonging to the civic category, whereas Damian placed 70% of the mission statements in that category. The diagonal values of Table 1 represent ratings on which the two raters agreed exactly. Thus, the raters agreed on their assignment of 5% percent of the mission statements to the social category, 3% percent to the emotional category, 9% to the cognitive category, and 54% to the civic category. Thus, from a simple percentage agreement perspective, the two raters agreed on 71% of the ratings they assigned. The percent agreement calculation can be derived by summing the values found in the diagonals (i.e., the proportion of times that the two raters agreed). Note that the resultant value of 71% generally represents good agreement:

$$P_A = .05 + .03 + .09 + .54 = .71.$$

Yet the high percentage agreement statistic is somewhat artificially inflated given that more than half of the school mission statements were rated as having a civic theme. Consequently, a rater with no knowledge or training could actually simply assign a mission statement to the civic category when in doubt, and the raters would end up with percentage agreement statistics that look very good simply because most schools had a civic purpose as their dominant

Table 1 Example Data Matrix: Rater Classifications of 100 School Mission Statements

		Steve				Marginal Total
		Social	Emotional	Cognitive	Civic	
Damian	Social	.05 (.01)	0 (0)	0 (.01)	0 (.03)	.05
	Emotional	.01 (.01)	.03 (0)	0 (.01)	.01 (.03)	.05
	Cognitive	.04 (.04)	.01 (.02)	.09 (.02)	.06 (.12)	.20
	Civic	.10 (.14)	.03 (.05)	.03 (.08)	.54 (.43)	.70
	Marginal total	.20	.07	.12	.61	1.00

Note: Values in parentheses represent the expected proportions on the basis of chance associations, i.e., the joint probabilities of the marginal proportions.

theme. Unfortunately, such an artificially inflated agreement statistic deceives us into believing that the two raters are perhaps more adept at coding the statements than they actually are. The raters actually agree less than half of the time (44% to be exact) when they are assigning codes to mission statements that they rate as falling into all categories other than the civic category.

To correct for the problem of inflation and to provide a more accurate estimate of rater agreement, we can calculate Cohen's kappa. To calculate kappa, we must begin by multiplying the marginal totals in order to arrive at an expected proportion for each cell (reported in parentheses in the table). Summing the product of the marginal values in the diagonal, we find that on the basis of chance alone, we expect an observed agreement value of .46:

$$P_c = .01 + 0 + .02 + .43 = .46.$$

Kappa provides an adjustment for this chance agreement factor. Thus, for the data in Table 1, kappa would be calculated as

$$\kappa = \frac{.71 - .46}{1.0 - .46} = .46.$$

In practice, kappa may be interpreted as the proportion of agreement between raters after accounting for chance. Consequently, a kappa value of zero suggests that the two raters agreed no more frequently than we would predict on the basis of chance alone. Furthermore, kappa can actually take on negative values if the raters were to agree less frequently than we would predict by chance alone, given the marginal values. Benchmarks for interpreting kappa are suggested in Table 2.

Three major assumptions underlie the use of Cohen's kappa:

1. Each of the units of analysis is independent.
2. Categories of the nominal scale must be mutually exclusive and exhaustive.
3. Raters must not work together to arrive at their final ratings.

Table 2 Benchmarks for Interpreting Kappa

<i>Kappa Statistic</i>	<i>Strength of Agreement</i>
<0.00	Poor
0.00 0.20	Slight
0.21 0.40	Fair
0.41 0.60	Moderate
0.61 0.80	Substantial
0.81 1.00	Almost perfect

Source: From Landis & Koch, 1977 (p. 165).

Using the Computer

Cohen's kappa may be calculated in SPSS by using the crosstabs procedure. The crosstabs procedure produces the output shown in Table 3.

Table 3 SPSS Crosstabs Output Calculating Cohen's Kappa

	<i>Symmetric Measure</i>			
	<i>Value</i>	<i>Asymp. Std. Error^a</i>	<i>Approx. T^b</i>	<i>Approx. Sig.</i>
Measure of Agreement				
Kappa	.458	.076	7.392	.000
N of Valid Cases	100			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

—Steven E. Stemler

See also Interrater Reliability

Further Reading

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- Kvalseth, T. O. (1989). Note on Cohen's kappa. *Psychological Reports*, 65, 223–226.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved May 10, 2006, from <http://pareonline.net/getvn.asp?v=9&n=4>

Uebersax, J. (2002). *Statistical methods for rater agreement*. Retrieved August 9, 2002, from <http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Hall, D. G., Veltkamp, B. C., & Turkel, W. J. (2004). Children's and adults' understanding of proper namable things. *First Language*, 24(1), 5–32.

D. Geoffrey Hall and his colleagues explored 5-year-olds' and adults' beliefs about entities that receive reference by proper names. In Study 1, two tasks were used. The first was a listing task, in which participants stated what things in the world can and cannot receive proper names. The second was an explanation task, in which they explained why some things merit proper names where two independent coders reached agreement which was verified by **Cohen's kappa**. Children's lists of proper namable things were more centered than adults' on living animate entities and their surrogates (e.g., dolls and stuffed animals). Both children's and adults' lists of nonnamable things contained a predominance of artifacts. Both age groups offered similar explanations for proper namability, the most common of which pertained to the desire or need to identify objects as individuals (or to distinguish them from other objects). In Study 2, the main results of the Study 1 listing task were replicated, using a modified set of instructions. The findings establish a set of norms about the scope and coherence of children's and adults' concept of a proper namable entity, and they place constraints on an account of how children learn proper names.

COMPLETE INDEPENDENCE HYPOTHESIS

Most studies involve multiple observations on multiple variables. With k variables, there are $k(k - 1)/2$ bivariate correlations among the measures. Each of

the individual correlations could be evaluated for statistical significance. Additionally, a multiple correlation could be obtained by regressing each of the k variables on the remaining $k - 1$ variables. Due to the large number of parameters, trying to ascertain the joint significance of the entire set of correlations is complex. A more direct approach is to evaluate the k by k symmetric correlation matrix R for complete independence. A population in which the null hypothesis of complete independence is true is characterized by a population correlation matrix $P = I$, the identity matrix, in which all correlations are equal to 0. If this null hypothesis can be rejected, it may be concluded that the variables in the data set are significantly related. Two common statistical tests for assessing complete independence are denoted L_1 and L_2 . L_1 is based on Fisher's Z or $\tanh^{-1}$ transformation of r_{ij} , the bivariate correlation between variables i and j . Because $\tanh^{-1}(r) = \{\log(1 + r) - \log(1 - r)\}/2$, where $\log$ is the natural or Naperian log, has variance of $1/(N - 3)$, the statistic L_1 is distributed as a chi-square with $k(k - 1)/2$ degrees of freedom where k is the number of variables:

$$L_1 = (N - 3) \sum_{i=1}^k \sum_{j=1}^k \tanh^{-1}(r_{ij})^2 \quad i < j.$$

L_2 is based on the distribution of the determinant of R , denoted $|R|$, which ranges between 0 and 1. Values closer to 0 indicate greater dependence among the measures, and values closer to 1 indicate greater independence. The statistic L_2 is also distributed as a chi-square with $k(k - 1)/2$ degrees of freedom for k variables, where the multiplier $\rho = -(N - 1 - (2k + 5)/6)$ and N is the sample size:

$$L_2 = -\rho \log |R|.$$

L_1 and L_2 have both been subjected to Monte Carlo sampling studies to evaluate Type I error rates (i.e., incorrectly rejecting the null hypothesis when it is true). A hypothesis test is considered to be biased if the estimated Type I error rate exceeds the test size α . When N is small relative to the number of variables, L_2 does not perform well. L_2 is biased when N is less

than 4 times the number of variables. Even by employing finite sample correction terms, L_2 is biased when N is less than twice the number of variables. In contrast, L_1 is an unbiased hypothesis test, even for small N .

Power comparisons between L_1 and L_2 indicate that L_1 is a more powerful test of $P = I$ than is L_2 . In small samples, relative to the number of variables, L_1 should be preferred in terms of both Type I and Type II error rates.

—John R. Reddon and James S. Ho

See also Correlation Coefficient; Type I Error; Type II Error

Further Reading

Reddon, J. R. (1990). The rejection of the hypothesis of complete independence prior to conducting a factor analysis. *Multivariate Experimental Clinical Research*, 9, 123–129.

COMPLETION ITEMS

Educational and psychological tests are composed of test items (questions or statements requiring a response) in many formats. One common format is the completion item. In this context, completion items come in several forms with several names.

Most commonly, completion items include some form of constructed response and are frequently called constructed-response items. The item does not contain options from which a person could select a response but requires the individual to construct a response. This may be accomplished by asking a complete question or providing a statement that must be completed, as the following examples demonstrate.

Item 1: What type of reliability can be estimated from one form of a test administered on a single occasion?
Answer: Coefficient alpha, split-half reliability.

Item 2: Describe one advantage of true-false items compared with multiple-choice items. Answer: Less testing time per item, easier to construct.

Item 3: To be most useful, norms should be representative, relevant, and _____. Answer: recent.

Some people distinguish between constructed-response items and completion items. Many people consider completion items to be primarily of the short-answer type. Constructed-response items, on the other hand, may also include extended response and essay items (requiring extensive responses, potentially several paragraphs long), configural response items (such as items requiring manipulation of schematic diagrams), or computation problems (commonly found in mathematics, where the individual must compute the answer).

The use of completion items, including constructed-response items, has advantages and disadvantages. Among the advantages, completion items are appropriate when the objective being measured requires a written response; they are relatively easy to construct and, when responses are short or composed of a single word, easy to score; short-answer items can assess higher-order thinking skills; and completion items allow for novel responses or solutions. Disadvantages include scoring difficulty because of the many possible correct responses, which may reduce reliability of scores; need for longer testing time, compared with multiple-choice testing, to achieve adequate reliability; low likelihood of assessing higher-order thinking skills with single-word answer formats; and because constructed-response items take more time to complete, limitation of the content that can be covered in a single test period.

Although it is possible to construct multiple-choice items to measure higher-order thinking skills, it appears that the range of cognitive skills addressed by completion items is larger than the range addressed by multiple-choice testing. Empirical evidence suggests that when written to tap the same content and cognitive skill, completion items and multiple-choice items measure the same construct; when written to tap different cognitive skills, the two formats appear to measure substantially different constructs. So it is not the format that determines what is being measured, but the nature and quality of the problem presented by the item, whatever its format.

—Michael C. Rodriguez

See also Essay Items; Multiple Choice Items

Further Reading

- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.
- Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah, NJ: Erlbaum.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Bornstein, M. H., Hahn, C.-S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, 24(3), 267–304.

Children participated in four longitudinal studies of specific and general language performance cumulatively from age 1 year and 1 month to age 6 years and 10 months. Data were drawn from age-appropriate maternal questionnaires, maternal interviews, teacher reports, **completion items**, experimenter assessments, and transcripts of children's own spontaneous speech. Language performance at each age and stability of individual differences across age in girls and boys were assessed separately and together. Across age, including the important transition from preschool to school, across multiple tests at each age, and across multiple reporters, children showed moderate to strong stability of individual differences; girls and boys alike were stable. In the second through fifth years, but not before or after, girls consistently outperformed boys in multiple specific and general measures of language.

COMPUTATIONAL STATISTICS

The term *computational statistics* has two distinct but related meanings. An older meaning is synonymous with the term *statistical computing*, or simply computations for use in statistics. The more recent meaning emphasizes the *extensive* use made of computations in statistical analysis.

Statistical Computing: Numerical Analysis for Applications in Statistics

Statistical analysis requires computing, and applications in statistics have motivated many of the

advances in numerical analysis. Particularly noteworthy among the subareas of numerical analysis are numerical linear algebra, numerical optimization, and the evaluation of special functions. Regression analysis, which is one of the most common statistical methods, as well as other methods involving linear models and multivariate analysis, requires fast and accurate algorithms for linear algebra. Linear regression analysis involves analysis of a linear model of the form $y = Xb + e$, where y is a vector of observed data, X is a matrix of observed data, b is a vector of unknown constants, and e is an unobservable vector of random variables with zero mean. Estimation of the unknown b is often performed by minimizing some function of the residuals $r(b) = y - Xb$ with respect to b . Depending on the function, this problem may be a very difficult optimization problem.

Numerical Linear Algebra

A very common approach to linear regression analysis is to minimize the sum of the squares of the residuals. In this case, the optimization problem reduces to a linear problem: Solve $X^T Xb = X^T y$. (Here, the superscript T means transpose.) This problem and other similar ones in the analysis of linear models are often best solved by decomposing X into the product of an orthogonal matrix and an upper triangular matrix without ever forming $X^T X$. Methods for doing this have motivated much research in numerical linear algebra.

Other important applications of numerical linear algebra arise in such areas as principal components analysis, where the primary numerical method is the extraction of eigenvalues and eigenvectors.

Numerical Optimization

Many statistical methods, such as regression analysis mentioned above, are optimization problems. Some problems, such as linear least squares, can be formulated as solutions to linear systems, and then the problems fall into the domain of numerical linear algebra. Others, such as nonlinear least squares and many maximum likelihood estimation problems, do not have closed-form solutions and must be solved by iterative methods, such as Newton's method,

quasi-Newton methods, general descent methods such as Nelder-Mead, or stochastic methods such as simulated annealing.

Another class of optimization problems includes those with constraints. Restricted maximum likelihood and constrained least squares are examples of statistical methods that require constrained optimization.

Evaluation of Special Functions

Methods for evaluation of cumulative distribution functions (probabilities) and inverse cumulative distribution functions (quantiles) are important in all areas of applied statistics. Some evaluations are straightforward, such as for z scores or for p values of common distributions such as t or F , but others are much more complicated. Computations involving posterior distributions in Bayesian analyses are often particularly difficult. Most of these computations are performed using Markov chain Monte Carlo methods.

Random Number Generation

Monte Carlo methods are widely used in statistics, both in development of statistical methodology and in applications of statistical methods. Monte Carlo methods require good programs for generating random numbers, firstly from a uniform distribution and secondly from various other distributions. (Standard methods for generating random variates from any given distribution utilize transformations of random variates from a uniform distribution.)

Most random number generators are cyclic; that is, they repeat after some fixed period. If the period is long enough, this repetition is not a problem, but many of the widely used random number generators have a period of approximately 2^{31} . This is much too small for serious work in Monte Carlo simulation. There are several good generators with periods greater than 2^{100} , and this should be a minimal standard for important Monte Carlo work.

In addition to problems with the period of generators, many generators have serious deficiencies in regard to the “randomness” of their output. Testing random number generators is a difficult task because of the nature of the problem: There is no standard

“answer” with which to compare the output of the generator. The ways in which a generator can produce unacceptable results are many and varied. Some generators are thought to be good until someone discovers a systematic departure from randomness, sometimes several years after the generator entered service. Anyone using Monte Carlo methods should be very careful to use random number generators with long periods and with no known departures from randomness. Likewise, statistical software developers should remain abreast of current research in the area so as to be able to provide high-quality random number generators.

Computational Statistics: Computationally Intensive Statistical Methods

Many statistical methods require extensive computations, not just because the data set is large, but because the method itself involves simulation of a statistical distribution or because the method requires multiple analyses.

Resampling and Data-Partitioning Methods

An effective approach to data analysis is to use the empirical cumulative distribution function to make inferences about the distribution of the observed data. In this approach, the underlying, unknown distribution is approximated by a discrete uniform distribution with mass points at the values of the observed data. This empirical distribution of the sample is then used to make statistical inferences about the unknown distribution of the population. This is called a *bootstrap method*. Often in a bootstrap method, the sample is resampled randomly. This approach can be useful in reducing bias of statistical procedures or for estimating variances or setting confidence intervals.

Related methods involve partitioning the sample, or analyzing subsets of the sample and then combining the results. This kind of data partitioning includes so-called *jackknife methods*, which can be used to reduce bias or to estimate variance, and *cross-validation methods*, which can be used to choose between statistical models and parameter estimates.

Statistical Inference Based on Monte Carlo Methods

Many statistical methods require simulation, either by randomly resampling the given data as mentioned above or by generating random data under the assumptions of a hypothesis to be studied. In a Monte Carlo statistical test using a given sample, multiple data sets are generated according to the null hypothesis, a test statistic is computed from each, and then the test statistic from the given sample is compared to the set of test statistics from the simulated data sets. If the observed test statistic is extreme within this set, the hypothesis is rejected.

Discovery of Structure in Data and Statistical Learning

In many cases, there is no obvious model for analysis of a given set of data. The data may have been collected for one purpose, possibly just business record-keeping, and then it may be used as a source of information about any number of new questions, some of which are not even enunciated clearly. This kind of exploratory analysis is sometimes called *data mining* or *knowledge discovery*. The main objective is to discover relationships or structure in the data that was perhaps not anticipated and then to interpret these relationships in meaningful ways.

Graphical displays from multiple perspectives are important in these exploratory analyses.

Software

Good and easy-to-use software is very important, both in statistical analysis of given data and in development of new statistical methodology.

A wide range of software is available for different purposes. Many simple analyses can be performed using a spreadsheet program, such as Excel or Lotus. Other analyses require more powerful software that provides a wider array of analyses. Software packages such as SAS, SPSS, and Minitab implement the standard statistical analyses as well as many more-specialized analyses, all in an integrated environment

that provides extensive abilities for data management and for graphical display of the data.

Some specialized analyses have a limited range of applications. Often the natural data structures for these applications are different from the more standard data structures for other statistical data. For such analyses and applications, stand-alone statistical software packages are available.

For implementing new research methods and for many exploratory analyses, an integrated programming environment is useful. Software packages such as SAS/IML and S-Plus provide flexible program-control structures, as well as a large library of standard functions.

The open source movement is important in the development of statistical software. Statisticians have traditionally shared programs with one another, but now there is a large, integrated statistical software system, called R, that benefits from input from statisticians around the world. The package is freely distributed (with restrictions on redistribution), and the source code is available for anyone to inspect and modify.

—James E. Gentle

See also Data Analysis Toolpak; Eigenvalues; Markov Chain Monte Carlo Methods; Monte Carlo Methods

Further Reading

- Gentle, J. E. (2002). *Elements of computational statistics*. New York: Springer.
- Gentle, J. E., Härdle, W., & Mori, Y. (Eds.). (2004). *Handbook of computational statistics: Concepts and methods*. New York: Springer.
- Lange, K. (1999). *Numerical analysis for statisticians*. New York: Springer.
- Thisted, R. A. (1988). *Elements of statistical computing*. New York: Chapman & Hall.

COMPUTERIZED ADAPTIVE TESTING

Computerized adaptive testing (CAT) is a method of administering tests that adapts to the examinee's trait level. A CAT test differs profoundly from a

paper-and-pencil test. In the former, different examinees are tested with different sets of items. In the latter, all examinees are tested with an identical set of items. The major goal of CAT is to fit each examinee's trait level precisely by selecting test items sequentially from an item pool according to the current performance of an examinee. In other words, the test is tailored to each examinee's θ level, so that able examinees can avoid responding to too many easy items, and less able examinees can avoid being exposed to too many difficult items. The major advantage of CAT is that it provides more-efficient latent trait estimates (θ) with fewer items than would be required in conventional tests.

The earliest large-scale application of CAT is the computerized version of the Armed Services Vocational Aptitude Battery (ASVAB), now administered to more than half a million applicants each year. The paper-and-pencil version of the ASVAB takes 3 hours to complete, and the CAT version takes about 90 minutes. With the CAT version, an examinee's qualifying scores can be matched immediately with requirements for all available positions. CAT has become a popular mode of assessment in the United States. In addition to the ASVAB, examples of large-scale CATs include the Graduate Record Examinations (GRE), the Graduate Management Admission Test, and the National Council of State Boards of Nursing. The implementation of CAT has led to many advantages, such as new question formats, new types of skills that can be measured, easier and faster data analysis, and faster score reporting. Today the CAT GRE is administered year-round, which allows examinees to choose their own date and time for taking it, whereas the paper-and-pencil version is administered only 3 times per year.

Item Selection in CAT

The most important ingredient in CAT is the item selection procedure, which selects items during the course of the test. According to M. F. Lord, an examinee is measured most effectively when test items are neither too difficult nor too easy. Heuristically, if the examinee answers an item correctly, the next item selected should be more difficult; if the answer is

incorrect, the next item should be easier. Because different examinees receive different tests, in order to equate scores across different sets of items, it is necessary to use a convenient probability model for item responses, and this can be achieved by item response theory (IRT). According to IRT modeling, a difficult item will have large b -value, and an easy item will have small b -value. Knowing the difficulty levels of all the items in the pool, one can possibly develop an item selection algorithm based on branching. For instance, if an examinee answers an item incorrectly, the next item to be selected should have a lower b -value; if the examinee answers correctly, the next item should have a higher b -value.

In 1970, Lord proposed an item selection algorithm as an extension of the Robbins-Monro process, which has been widely used in many other areas, including engineering control and biomedical science. The Robbins-Monro process has been proved a method in minimizing the number of animals required to estimate the acute toxicity of a chemical. In order to use the method in CAT, item difficulty levels for all the items in the item pool must be calibrated before testing. Let $b_1, b_2, \dots, b_n$ be a sequence of the difficulty parameters after administering n items to the examinee. The new items should be selected such that b_n approaches a constant b_0 (as n are indefinitely large), where b_0 represents the difficulty level of an item that the examinee has about a 50% chance of answering correctly, or $P\{X_n = 1 \mid \theta = b_0\} \approx 1/2$. Because our goal is to estimate θ , knowing b_0 , we can use b_0 as a reasonable guess for θ . Notice that b_0 can be linearly transformed to any meaningful score scale, which makes it convenient for us to score the examinee's test responses by a function of b_0 . Lord, writing in 1970, proposed several rules based on the Robbins-Monro process and envisioned that such testing could be implemented when computers became sufficiently powerful. A specific example of the item selection rule can be described by the following equation:

$$b_{n+1} = b_n + \frac{d_1}{n}(x_n - 0.5),$$

where x_n is the item response on the n th item ($x_n = 1$ if the answer is correct, $x_n = 0$ if the answer is incorrect),

and d_1 is a positive number chosen before the testing. Clearly, the motivation for the adaptive design is to tailor the difficulty levels of the items administered to the latent trait of the examinee being tested.

In 1980, Lord proposed the maximum information (MI) approach, which has become a standard item selection procedure. Let $\hat{\theta}_n$ be an estimator of θ based on n responses. The MI method selects the item with the maximum Fisher item information evaluated at $\hat{\theta}_n$ as the next item. Under IRT, maximizing Fisher information means intuitively matching item difficulty parameter values with the latent trait level of an examinee. In addition, items with high discrimination, or equivalently, high a -parameter value, will be preferentially selected by the algorithm. Maximizing the Fisher information will lead to minimizing the sample variance of $\hat{\theta}_n$, and that makes $\hat{\theta}_n$ the most efficient. For this reason, the MI method has become the most popular item selection method. Though CAT was originally developed in educational assessments, it should be effective in many other areas, such as cognitive diagnosis and quality of life assessment.

Nonstatistical Constraints in CAT Design

In order to design an operational CAT test, the set of items selected for each examinee must satisfy certain nonstatistical constraints, such as item exposure control and content balance. The more constraints one has to impose, the fewer degrees of freedom one can include in a design. To design a good CAT algorithm, many complex controls are needed.

The item exposure rate for each item is defined as the ratio of the number of times the item is administered to the total number of examinees. Since CAT is designed to select the best items for each examinee, certain types of items tend to be always selected by the computers, and many items are not selected at all, thereby making item exposure rates quite uneven. Because CAT tests are usually administered to small groups of examinees at frequent time intervals, examinees who take tests earlier may share information with those who will take tests later,

escalating the risk that many items may become known. Therefore, item exposure rates must be controlled. A number of methods have been developed to control item exposure rate. The most common method of controlling exposure rate was developed by Simpson and Hetter, whose general idea is to put a “filter” between selection and administration such that an item that is selected by the MI criterion is evaluated to determine whether it will be administered. In this way, the exposure rate can be kept within a certain prescribed value. The Simpson and Hetter approach suppresses the use of the most over-exposed items, usually items with high a -parameters, and spreads their use over the next tier of over-exposed items. Chang and Ying proposed the a -stratified method with the objective of limiting the exposure of any given item by using that item at the most advantageous point in testing. It attempts to control item exposure by using less-discriminating items early in the test, when θ estimation is least precise, and saving highly discriminating items until later stages, when finer gradations of θ estimation are required. One of the advantages of the a -stratified method is that it tends to equalize the item exposure rates for all the items in the pool.

Methods have been developed to handle various types of content-balancing constraints. In particular, linear programming (LP) has been used to handle flexible content balancing, which selects items using LP based on numerous simultaneous constraints involving statistical and content considerations. One constraint is to maximize item information. Other constraints can be mathematical representations of the test specifications or a model to control for item overlap. A weighted deviations model has been proposed that includes LP as one component and also incorporates some heuristic steps when LP's solution may not be suitable. Actually, both the weighted deviations model and LP methods are capable of dealing with multiple constraints, among them content balancing constraints, exposure control, statistical optimization, and others. They are in general very powerful, with relatively more intensive computation than other methods. Recently, such content balancing constraints as item pool stratification have been proposed, and

this method has been generalized for flexible content balancing.

Issues to Be Addressed

Although CAT has many advantages, issues regarding large-scale applications need to be addressed. One of them is the compatibility between CAT and paper-and-pencil tests. It has been speculated that some examinees may get much lower scores on a CAT test than they would on a paper-and-pencil version. As evidence of this, the Educational Testing Service (ETS) found that the GRE CAT system did not produce reliable scores for a few thousand examinees in 2000. ETS offered them a chance to retake the test at no charge. Another vital issue is test security and large-scale item theft. In August 2002, ETS suspended the CAT GRE and reintroduced paper-and-pencil-based versions in China, Hong Kong, Taiwan, and Korea (www.ets.org, August 20, 2002) following an investigation that uncovered a number of Web sites offering questions from live versions of the CAT GRE. Another issue is item pool usage. An examination of item usage within the GRE CAT pools found that as few as 12% of the available items can account for as many as 50% of the items actually administered. Without effective remedial measures, this state of affairs could significantly undermine the future of CAT.

In response to the problems that emerged from the initial large-scale applications, researchers proposed corrective procedures. They focused principally on refinement of item selection methods and also on how to assess the severity of organized item theft activities. The underestimation-of-performance problem of the GRE is very likely caused by the item selection strategy, which heavily relies on the items with the highest discrimination at the beginning of the test. A reasonable solution is to use a weighting mechanism to estimate the weights of the likelihood function during the early-stage estimation. To stabilize the initial estimation of the examinee's latent trait, items with low discrimination, instead of those with high discrimination, may be used at the beginning of the test.

To improve test security, several theorems have been derived on the basis of the hypergeometric

distribution family for addressing questions such as "in order to compromise 200 items from a given item pool, how many thieves, at the most, would need to take the test?" The results may shed light on the relationship between optimal item pool size and test security.

It is important to be aware that test security can be enhanced by evenly using all the items in a pool. A computer using the constrained MI item selection method will not select many items in the pool, so the actual pool size becomes much smaller than the original pool, which makes item theft much easier. On the other hand, if the computer algorithm selects only high- a items, we may have to force item writers to generate only high- a items. Item writers may control such characteristics as item content and item difficulty level, but it is extremely challenging to produce only highly discriminating items. The common practice for generating more relatively high- a items is to discard items whose a -parameter values are lower than a given threshold. Once items are included in the pool, they have already undergone rigorous review processes and shown no problems. Items with relatively lower discrimination parameters are still of good quality and should be used. Obviously, test security can be greatly enhanced by increasing the use of lower- a items.

This research indicates that structuring an operational CAT exam with only several hundred items should be considered a design flaw. A high-stakes CAT exam must have a large item pool. This can be accomplished partly by including many items that have never been selected by the current item selection algorithms. Therefore, test security can be significantly improved by increasing the pool size and by evenly selecting all the items in the pool. If the item pool is sufficiently large, an examinee who has studied compromised items has relatively little advantage. But if the pool is small, the advantage can be huge.

Despite its limitations, CAT undoubtedly has a great future because cutting-edge developments in technology will enable us to solve the problems encountered in current large-scale applications.

—Hua-Hua Chang and Zhiliang Ying

See also Armed Services Vocational Aptitude Battery; Graduate Record Examinations

Further Reading

- Chang, H., & Ying, Z. (1999). a-Stratified computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.
- Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67, 387–398.
- Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper & Row.
- Lord, M. F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22, 400–407.
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270.
- Yi, Q., & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56, 359–378.

**COMREY, ANDREW L.
(1923–)**

Andrew L. Comrey was born in Charleston, West Virginia, on April 14, 1923. His childhood was marred by the Great Depression of the 1930s; however, he was a brilliant student and entered Union College with a full scholarship. While he was at Union College, his psychology teacher, Ernest M. Ligon, introduced him to psychological testing. Comrey worked in Ligon's laboratory giving Stanford-Binet IQ tests to young people in the Character Research Project. During this time, he read Louis Thurstone's *Vectors of Mind* and planned to attend the University of Chicago to study with Thurstone. However, World War II changed his plans. Andrew completed his BS degree in science and entered the U.S. Navy. During his service, he met and married Barbara Sherman, who was also serving in the military. They have two daughters, Cynthia and Corinne.

After the war, Comrey attended the University of Southern California. He studied measurement, psychometrics, and statistics with J. P. Guilford and earned his PhD in 1949. His dissertation on fundamental measurement included a treatise on a method

of absolute ratio scaling. This method, published in *Psychometrika*, was among the most cited studies on scaling at that time.

Comrey's first academic appointment was at the University of Illinois. In 1951, he accepted a faculty position at the University of California, Los Angeles (UCLA). He has been at UCLA since 1951. During his time at UCLA, he was a Fulbright Research Fellow and held a National Science Foundation senior postdoctoral research fellowship. He has served as president of the Society for Multivariate Experimental Psychology. His major research contributions included the development of his own complete system of factor analysis. He invented the minimum residual method of factor extraction and the tandem criteria of rotation.

The minimum residual method of factor extraction is a controversial method that avoided the problem associated with communality estimates. The tandem criteria involved a two-phase procedure to obtain a simple structure solution. The first of the two criteria is very useful in finding general factors.

Comrey was the first researcher to write a fully integrated computer program that would process raw data through correlations, factor extraction, and rotation. He later used this process to develop the Comrey Personality Scales (CPS). The CPS was the result of his research on the Minnesota Multiphasic Personality Inventory and other personality tests. During the development of the CPS, he created the factored homogeneous item dimension as a basic unit of analysis in factor analysis.

Comrey has published more than 150 articles, chapters, and books. His textbook *A First Course in Factor Analysis* remains a popular and highly regarded work. It has been translated into Japanese and Italian.

—Howard B. Lee

Further Reading

- Comrey, A. L. (1976). Mental testing and the logic of measurement. In W. L. Barnette (Ed.), *Readings in psychological tests and measurement*. Baltimore: Williams & Wilkins.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Andrew L. Comrey: <http://www.today.ucla.edu/2002/020312comrey.html>

COMREY PERSONALITY SCALES

The Comrey Personality Scales (CPS), developed by factor analysis, is a personality inventory of 180 multiple-choice items. Each item uses one of two possible 7-choice answer scales. Scale X has the following possible answers: 7 (*always*), 6 (*very frequently*), 5 (*frequently*), 4 (*occasionally*), 3 (*rarely*), 2 (*very rarely*), and 1 (*never*). Scale Y has the following possible answers: 7 (*definitely*), 6 (*very probably*), 5 (*probably*), 4 (*possibly*), 3 (*probably not*), 2 (*very probably not*), and 1 (*definitely not*). A sample item using the X scale is "I love to work long hours." The CPS is published by the Educational and Industrial Testing Service (www.edits.net).

The CPS measures eight major factors of personality that were identified through a unique set of factor analytic methods and procedures. First, a collection of factored homogeneous item dimensions (FHIDs) was developed by culling ideas for important personality variables from many existing personality inventories and other sources. Then, each personality concept was defined in clear terms. Next, multiple-choice items were written that were deemed possible measures of each construct. Seven-choice items were chosen because seven proved to be optimal for giving good item response distributions and reliable items while not demanding too much from a respondent. Each FHID consisted of two items that were positively stated and two items that were negatively stated with respect to the construct dimension. This process was carried out for a large number of potentially useful constructs.

Next, item factor analyses were conducted in which the variables were items from six or more possibly useful constructs. If a factor was identified by the items measuring a given hypothesized construct, and no other, the items identifying that factor were

considered an FHID and a way of measuring that dimension. Many factor analyses were conducted, and items and dimensions were refined, culled, discarded, and replaced until a substantial number of FHIDs were selected. In these analyses, great care was exercised to make sure that each FHID was conceptually distinct from each other FHID in the analysis.

The number of factors that emerged in analyses of conceptually distinct FHIDs was found to be strictly limited. In fact, the author's extensive empirical investigations succeeded in turning up only eight major factors. These eight factors constitute the personality taxonomy on which the CPS is based. They are as follows: trust versus defensiveness (T), orderliness versus lack of compulsion (O), social conformity versus rebelliousness (C), activity versus lack of energy (A), emotional stability versus neuroticism (S), extraversion versus introversion (E), mental toughness versus sensitivity (M), and empathy versus egocentrism (P). Each of these major personality factors has been identified in previous studies under one name or another. CPS is unique because of the particular combination of factors that makes up its taxonomy and because each factor has been identified by a compelling rationale.

The CPS taxonomy has been validated with various kinds of samples in many different cultural settings. Factorial validity has been extensively documented for the eight CPS factors. These factors have been shown to be useful for predicting outcomes with respect to a number of different practical criteria. The clinical significance of CPS scores, particularly extreme scores, high and low, has been well documented. Brief descriptions and summaries of relevant published articles are given in the *Manual and Handbook of Interpretations* referenced below. A description of the factor analytic procedures used in developing the CPS is given in *A First Course in Factor Analysis*.

—Andrew L. Comrey

See also Minnesota Multiphasic Personality Inventory; NEO Personality Inventory; Personality Tests

Further Reading

Comrey, A. L. (1970). *Manual for the Comrey Personality Scales*. San Diego, CA: Educational and Industrial Testing Service.

Comrey, A. L. (1995). *Manual and handbook of interpretations for the Comrey Personality Scales*. San Diego, CA: EDITS Publishers.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

CONDITIONAL PROBABILITY

Conditional probability is a mathematical description of the likelihood that a particular event will take place given the occurrence of a particular precursor event. Thus the probability of occurrence of event A is *conditional* on the probability of the prior occurrence of event B. Conditional probability is expressed as $P(A|B)$, which is read as “the probability of A given B.” In order to calculate the conditional probability of an event, there must be a priori knowledge of both the probability that the first of the events will occur and the probability that both events will occur together. The basic probabilities of the events involved can be expressed by a Venn diagram, as shown in Figure 1.

In Figure 1, the expressions used are defined as follows:

$P(A)$ is the probability of event A occurring alone

$P(B)$ is the probability of event B occurring alone

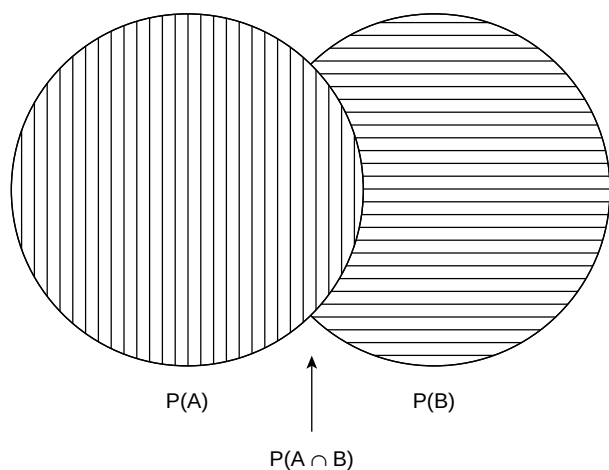


Figure 1 Venn Diagram Illustrating the Probabilities of Event A, Event B, and Events A and B Together

$P(A \cap B)$ is the probability of both events A and B occurring together, or the joint probability ($\cap$ is the mathematical symbol used to represent the intersection of two sets)

The conditional probability of an event can be calculated by dividing the probability of the occurrence of the first event into the joint probability. The mathematical formula for this calculation is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ or, similarly,}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

It follows that the probability of occurrence of the primary event (the denominator of the equation) cannot be equal to zero. Without this condition, the equation would be mathematically impossible, as well as illogical (one cannot estimate the probability of a particular outcome without the occurrence of the outcome's predecessor).

Conditional probability can be applied to many situations in a variety of fields. For example, it could be used to determine a person's chance of developing a particular disease given the disease's presence or absence in the person's familial background or the probability of a person using the word *cabbage* in a sentence, given that the word directly preceding it was *eggplant*. For gamblers, conditional probability could be used to compare the odds of winning a game; the chances of having a winning poker hand in five-card draw are significantly greater if a player is dealt three queens than if the player gets a pair of twos. In everyday life, your chances of being mugged in a city might be dependent on how often you walk alone at night, or your chances of heartburn might be conditional on your eating spicy foods.

As an example, suppose there is a 3% chance that a girl in a senior high school class is both a cheerleader and class president. Additionally, suppose that the chance of a girl in the senior class being a cheerleader is 8%. What are the chances of a girl being elected senior class president given that she is a cheerleader?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{.03}{.08} = .375 = 38\%$$

There is a 38% chance that a girl will be elected president of her senior class given that she is a cheerleader.

—Allison B. Kaufman

Further Reading

Glass, G. V., & Hopkins, K. D. (1995). *Statistical methods in education and psychology*. Boston: Allyn & Bacon.

Hays, W. L. (1994). *Statistics*. Belmont, CA: Wadsworth.

Conditional probability: http://en.wikipedia.org/wiki/Conditional_probability

CONFIDENCE INTERVALS

Confidence intervals (CIs) are common tools of inference, measuring how sure we are of our results. Confidence intervals do the following:

- Across studies, they tell us how accurately and consistently data operate over time.
- They invoke two primary concepts, intervals and confidence levels: Intervals are determined by the standard errors of statistics. Levels are chosen by the researcher and are given as percentages.

Simply put, a 95% confidence level says the method used by the researcher gives an interval that covers the true population parameter for 95% of the samples. For example, by calculating a confidence interval for your cholesterol level taken 20 times ($n = 20$), you can state how confident you are that the CI accurately contains your true cholesterol level. A range null hypothesis, say

160–200, is tested rather than a point null hypothesis (e.g., 180).

There exists a seesaw relationship between confidence levels and CIs: the higher the confidence level, the wider the interval or the larger the margin of error. The lower the confidence level, the narrower the interval or the smaller the margin of error. For the CI for the mean, the standard deviation also affects the margin of error, and there is more variance in the population if the interval is wider, as shown in Figures 1a and 1b. Figure 1c suggests that to make the margin of error smaller, the researcher must collect more data, which shrinks the margin of error because of the formula

$$\bar{X} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right),$$

where z^* is a z score related to the p value and is a measure of distance from the mean measured in standard deviations. The z^* for .05 is 1.96, equaling a 95% confidence level; z^* for .01 is 2.576, equaling a 99% confidence level.

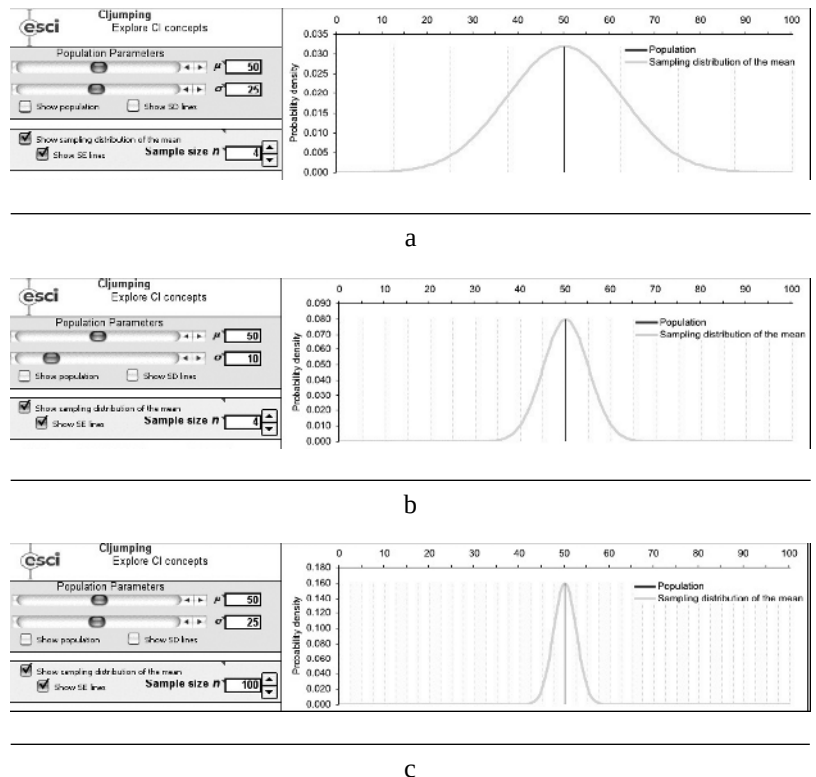


Figure 1 Effect of Changing Confidence Intervals

CIs should be used when reporting results for the following reasons:

- Graphical display of CI lends itself to enhanced understanding by readers.
- CIs are fairly easily obtained using common packages such as SPSS or the Exploratory Software for Confidence Intervals software developed by Cumming and Finch.
- CIs are helpful in compiling studies supporting meta-analytic thinking.

The American Psychological Association (APA) Task Force suggested that CIs should always be reported, and the APA *Publication Manual* said CIs were “the best” reporting device. One advantage of thoughtful use of CIs is that they provide a *graphical* tool to integrate or synthesize results across studies, thereby enhancing replicability. Researchers should present effect sizes as CIs because CIs contain much more information than significance tests.

—Mary Margaret Capraro

See also Hypothesis Testing; Significance Level

Further Reading

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532–575.

Confidence intervals, Rice Virtual Lab in Statistics page: http://davidmlane.com/hyperstat/confidence_intervals.html
Exploratory Software for Confidence Intervals: <http://www.latrobe.edu.au/psy/esci/>

CONSTRUCT VALIDITY

Validity refers to the degree that a test measures what it purports to measure. In terms of classical measurement theory, it is inappropriate to refer to the validity of a test. Instead, the *use* of the test is validated for a specific measurement purpose.

Instrument reliability, the consistency with which a test measures whatever it measures, is a necessary but insufficient condition in determining whether a test is

valid for a specific use. A test may be reliable but not valid for a particular purpose. However, a test that is not reliable cannot be valid for any meaningful measurement purpose.

In 1986, Crocker and Algina defined a construct as “informed scientific imagination.” A construct is a fiction that is used to explain reality. For example, intelligence, reading readiness, and self-determination are constructs used to study and communicate inferable educational and psychological phenomena.

There are a variety of ways to collect evidence of the validity of a test to measure a construct. Frequently, test publishers rely on the Pearson product-moment correlation as evidence of construct validity. A random sample of examinees may be given two different intelligence tests. The presence of a high correlation of a newly constructed intelligence test with an established intelligence test is cited as evidence of construct validity. However, this technique is no stronger than the external evidence that supports the established intelligence test as a measure of intelligence, which is oftentimes historically problematic.

Another technique is to postulate differential effects among groups. An experimental design is carried out on a known intervention to determine whether the outcome is aligned with the *a priori* differentiation. This method is also problematic for a variety of reasons, such as limitations on the reliability of the data-gathering instruments and the unexpected failure of the intervention.

An enigmatic method that is nevertheless frequently used is factor analysis. In exploratory factor analysis, a reduced set of underlying variables is discovered that purports to account for the variation of the test items. This reduced set is known as the factor solution, which constitutes the construct being measured. Nevertheless, a plethora of choices make this approach untenable. They include the choice of eigenvalue minimum to extract (e.g., 1.0), *a priori* number of factors to extract, method of extraction (e.g., principal components, principal axis, maximum likelihood), and rotation method (e.g., varimax, equamax).

Confirmatory factor analysis, and structural equation modeling in general, presents an improvement on exploratory factor analysis in that the former provides a method for testing a theoretic measurement model and the goodness of fit of the data to

Table 1 Multitrait–Multimethod Matrix Data

		Method 1			Method 2			Method 3		
	Trait:	A	B	C	A	B	C	A	B	C
Method 1										
	A	(.95)								
	B	.28	(.86)							
	C	.58	.39	(.92)						
Method 2										
	A	<u>.86</u>	.32	.57	(.95)					
	B	.30	<u>.90</u>	.40	.39	(.76)				
	C	.52	.31	<u>.86</u>	.55	.26	(.84)			
Method 3										
	A	<u>.74</u>	.10	.43	<u>.64</u>	.17	.37	(.48)		
	B	.10	<u>.63</u>	.17	.22	<u>.67</u>	.19	.15	(.41)	
	C	.35	.16	<u>.52</u>	.31	.17	<u>.56</u>	.41	.30	(.58)

Source: Adapted from Mosher, 1968.

Note: Correlations in parentheses are reliability coefficients, bold italics indicate validity coefficients, underscore indicates heterotrait monomethod coefficients, and regular type indicates heterotrait heteromethod coefficients.

that model. Despite its promise, however, Pedhazur and Schmelkin noted the “large and potentially bewildering number of models” (p. 670) that arise in confirmatory factor analysis, in addition to the many and stringent underlying assumptions that must be met.

In 1959, Campbell and Fiske provided a rigorous design for determining construct validity called the multitrait-multimethod matrix. The construct is partialled into constituent traits, which are then measured in a variety of ways. In 1995, for example, Field, Hoffman, & Sawilowsky defined the construct *self-determination* for students as consisting of the constituent traits of (1) knowing yourself, (2) valuing yourself, (3) being able to plan, (4) being able to act, and (5) being able to learn from outcomes. A battery of five instruments was developed to measure these traits via differing methodologies: (1) assessment of knowledge or skills, (2) behavioral observation checklist, (3) assessment of affect and belief, (4) teacher perception, and (5) parent perception.

The multitrait-multimethod matrix consists of four levels of data. They are, hierarchically from the highest to the lowest, reliability coefficients (usually

placed in parentheses), validity coefficients (in bold italics), heterotrait monomethod coefficients (underscored), and heterotrait heteromethod coefficients. As an illustration, consider the multitrait-multimethod matrix data in Table 1.

Since the multitrait-multimethod matrix was developed in 1959, there have been many unsuccessful attempts to determine the appropriate method of statistical analysis for the data in the matrix. The reason for the lack of success is that the data in the diagonals and triangles often conflict. These methods are based on heuristic argument, analysis of variance models, nonparametric analogs to the analysis of variance models, and confirmatory factor analysis.

Campbell and Fiske stated that evidence of construct validity requires the data for the coefficients at the top level (i.e., reliability) to be as high as possible and somewhat higher than those at the second level (validity), which in turn should be higher than those on the third level (heterotrait monomethod), and so forth.

A quick, distribution-free method with easily remembered critical values was provided for this analysis by Sawilowsky in 2002:

Table 2 Obtaining Minimum, Median, and Maximum Values for the I Test for Construct Validity*Reliability coefficients*

Original data:

.95, .86, .92, .95, .76, .84, .48, .41, .58

Ranked data:

.41, .48, .58, .76, .84, .86, .92, .95, .95

Minimum, median, maximum

.41, .84, .95

Validity coefficients

Original data:

.86, .90, .86, .64, .67, .56

Ranked data:

.56, .64, .67, .86, .86, .90

Minimum, median, maximum

.56, .765, .90

Heterotrait monomethod coefficients

Original data:

.28, .58, .39, .39, .55, .26, .15, .41, .30

Ranked data:

.15, .26, .28, .30, .39, .39, .41, .55, .58

Minimum, median, maximum:

.15, .39, .58

Heterotrait heteromethod coefficients

Original data:

.32, .57, .30, .40, .52, .31, .10, .43,
.10, .17, .35, .16, .17, .37, .22, .19, .31, .17

Ranked data:

.10, .10, .16, .17, .17, .17, .19, .22, .30, .31, .31,
.32, .35, .37, .40, .43, .52, .57

Minimum, median, maximum:

.10, .305, .57

Note: I = number of inversions.

$$I = \sum_{i=1}^k \sum_{j=1}^{N-k} (x_i > x_{j+3}) + \sum_{i=4}^{2k} \sum_{j=4}^{N-k} (x_i > x_{j+3}) \\ + \sum_{i=7}^{3k} \sum_{j=7}^{N-k} (x_i > x_{j+3}) \\ \begin{cases} 1 & \text{if } (x_i > x_{j+3}) \\ 0 & \text{if } (x_i \leq x_{j+3}) \end{cases}$$

The null hypothesis is that the coefficients in the matrix are unordered. This is tested against the alternative hypothesis of an increasing trend from the lowest level (heterotrait heteromethod) to the highest level (reliability coefficients).

The test statistic, I, is the number of inversions (also known as U statistics). Consider the coefficients in Table 1. The data are ranked. Next, the minimum, median, and maximum values are determined, as indicated in Table 2. Then, count the number of inversions, beginning with the minimum value of the lowest level of the heterotrait heteromethod coefficients, as indicated in Table 3. For example, there are no inversions from the initial value of .10. The second value, .305, has one inversion (.15, which is the minimum value on the heterotrait monomethod level). The third value, .57, has three inversions (.15, .39, and .56).

In this example, $I = 10$. The critical values for $\alpha = 0.05$ and 0.01 are 10 and 14, respectively. Thus, in this example, the null hypothesis that the values are unordered is rejected in favor of the alternative hypothesis of an upward trend. This constitutes evidence of construct validity. A complete table of critical values, and the associated p values, may be found in Sawilowsky (2002).

—Shlomo S. Sawilowsky

Table 3 Test for Trend (Construct Validity)

Level	Minimum		Median		Maximum	
	Value	I	Value	I	Value	I
Reliability	.41	0	.84	0	.95	0
Validity	.56	1	.765	0	.90	2
H-M	.15	0	.39	0	.58	3
H-H	.10	0	.305	1	.57	3

Notes: Total Inversions = 10. Probability $[I \leq 10] = 0.00796807$.
H-M = heterotrait monomethod, H-H = heterotrait heteromethod.

See also Content Validity; Criterion Validity; Face Validity; Predictive Validity; Reliability Theory; Validity Theory

Further Reading

Field, S., Hoffman, A., & Sawilowsky, S. (1995). *Self-Determination Knowledge Scale, Form A and Form B*. Austin, TX: Pro-Ed.

- Mosher, D. L. (1968). Measurement of guilt by self-report inventories. *Journal of Consulting and Clinical Psychology*, 32, 690–695.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Sawilowsky, S. (2000). Psychometrics versus datametrics. *Educational and Psychological Measurement*, 60, 157–173.
- Sawilowsky, S. (2000). Reliability. *Educational and Psychological Measurement*, 60, 196–200.
- Sawilowsky, S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development*, 35, 78–88.

CONTENT VALIDITY

In educational and psychological testing, the term *validity* refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association, 1999, p. 9). From this definition, it can be deduced that (a) tests can be evaluated only with respect to one or more specific testing purposes, (b) validation involves confirming the inferences derived from test scores, not confirming the test itself, and (c) evaluating inferences derived from test scores involves gathering and analyzing several different types of evidence. In many cases, the evidence most critical to evaluating the usefulness and appropriateness of a test for a specific purpose is based on *content validity*.

Content validity refers to the degree to which a test appropriately represents the content domain it is intended to measure. When a test is judged to have high content validity, its content is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. For many testing purposes, such as determining whether students have mastered specific course material or determining whether licensure candidates have sufficient knowledge of the relevant profession, evidence of content validity provides the most compelling argument that the test scores are appropriate for inferring conclusions about examinees’ knowledge, skills, and

abilities. For this reason, content validity evidence is critically important in supporting the use of a test for a particular purpose. Thus, content validation is an important activity for test developers, and research on improved methods for evaluating test content continues to this day.

Characteristics of Content Validity

As the American Educational Research Association definition of validity implies, empirical evidence and cogent theory are needed to support the validity of inferences derived from test scores. Investigations of content validity involve both evaluating the theory underlying the test and gathering empirical evidence of how well the test represents the content domain it targets.

Content validity has at least four aspects: domain definition, domain representation, domain relevance, and appropriateness of the test construction process. *Domain definition* refers to the process used to operationally define the content domain tested. Defining the domain is typically accomplished by providing (a) detailed descriptions of the content areas and cognitive abilities the test is designed to measure, (b) test specifications that list the specific content “strands” (subareas), as well as the cognitive levels measured, and (c) specific content standards, curricular objectives, or abilities that are contained within the various content strands and cognitive levels. Evaluating domain definition involves acquiring external consensus that the operational definition underlying the test is commensurate with prevailing notions of the domain held by experts in the field (e.g., certified public accountants’ verifying that the test specifications for the Uniform CPA Exam reflect the major knowledge and skill domains necessary for safe and effective practice in the profession).

The next aspect of content validity is *domain representation*, which refers to the degree to which a test represents and adequately measures all facets of the intended content domain. To evaluate domain representation, inspection of all the items and tasks on a test must be undertaken. The critical task is to determine whether the items fully and sufficiently

represent the targeted domain. Studies of domain representation typically use subject matter experts to scrutinize test items and judge the degree to which they are congruent with the test specifications. Sometimes, as in the case of state-mandated testing in public schools, subject matter experts judge the extent to which test items are congruent with curriculum framework. These studies of domain representation have recently been classified within the realm of test *alignment* research.

Related to domain representation is *domain relevance*, which addresses the extent to which each item on a test is relevant to the domain tested. An item may be considered to measure an important aspect of a content domain, and so it would receive high ratings with respect to domain representation. However, if it were only tangentially related to the domain, it would receive low ratings with respect to relevance. For this reason, studies of content validity may ask subject matter experts to rate the degree to which each test item is relevant to specific aspects of the test specifications. The studies then aggregate those ratings within each content strand to determine domain representation. Taken together, study of domain representation and relevance can help evaluate (a) whether all important aspects of the content domain are measured by the test and (b) whether the test contains trivial or irrelevant content. As Messick described in 1989 in his seminal treatise on validity, “Tests are imperfect measures of constructs because they either leave out something that should be included . . . or else include something that should be left out, or both” (p. 34). A thorough study of content validity, prior to assembling tests, protects against these potential imperfections.

The fourth aspect of content validity, *appropriateness of the test development process*, refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material. The content validity of a test can be supported if strong quality control procedures are in place during test development and if there is a strong rationale for the specific item formats used on the test. Examples of quality control procedures that

support content validity include (a) reviews of test items by content experts to ensure their technical accuracy; (b) reviews of items by measurement experts to determine how well the items conform to standard principles of quality item writing; (c) sensitivity review of items and intact test forms to ensure the test is free of construct-irrelevant material that may offend, advantage, or disadvantage members of particular subgroups of examinees; (d) pilot-testing of items, followed by statistical item analyses to select the most appropriate items for operational use; and (e) analysis of differential item functioning, to flag items that may be disproportionately harder for some groups of examinees than for others.

Conducting a Content Validity Study

As briefly mentioned earlier, many studies of content validity require subject matter experts to review test specifications and items according to specific evaluation criteria. Thus, a content validity study typically involves gathering data on test quality from professionals with expertise in the content domain tested. Content validity studies differ according to the specific tasks presented to the experts and the types of data gathered. One example of a content validity study is to give content experts the test specifications and the test items and ask them to match each item to the content area, educational objective, or cognitive level that it measures. In another type of study, the experts are asked to rate the relevance of each test item to each of the areas, objectives, or levels measured by the test.

The data gathered from these studies can be summarized using simple descriptive statistics, such as the proportion of experts who classified an item as it was listed in the test specifications or the mean relevance ratings for an item across all areas tested. A “content validity index” can be computed for a test by averaging these statistics over all test items. More-sophisticated procedures for analyzing these data have also been proposed, including a newer procedure based on experts’ judgments regarding the similarity of skills being measured by pairs of test items. Other studies that provide evidence of content validity

include job analyses (referred to as practice analyses for licensure testing). Job analyses are often conducted to operationally define the content domain to be tested. Data gathered from such analyses can be used to derive weights (e.g., proportions of test items) for specific content areas as well as to defend the specific areas tested.

Content Validity: Past, Present, and Future

The origins of contemporary, large-scale educational and psychological tests can be traced to the early 20th century. As the stakes associated with these tests increased, the methods used to evaluate tests also increased. The concept of content validity and the process of content validation emerged to address the limitations of purely statistical (correlational) approaches to test validation that were common in the early part of the 20th century. Content validity quickly became a popular term endorsed by validity theorists and by the Joint Committee on Testing Standards of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. That popularity waned in the middle 1970s, when a unitary conceptualization of validity centered on construct validity was proposed. Proponents of this unitary conceptualization suggest using terms such as *content representativeness* in place of content validity because content validity focuses on the test itself rather than on inferences derived from test scores. This perspective was incorporated into the current version of the American Educational Research Association's *Standards for Educational and Psychological Testing*, which uses the phrase "evidence based on test content" in place of content validity. However, not all test specialists agree, and in educational testing, the attention paid to content validation is increasing at a staggering pace.

Regardless of debates over terminology, the fundamental characteristics of test quality encompassed by content validity (i.e., domain definition, domain representation, domain relevance, and appropriate test construction process) will remain important criteria

for evaluating tests for as long as tests are used to make inferences regarding individuals' knowledge, skills, and abilities. Clearly, for interpretations of test results to be valid, (a) the content of a test needs to be congruent with the testing purpose, and (b) the content areas to which an assessment is targeted need to be adequately represented. Thus, for many educational and psychological tests, content validity is prerequisite for valid score interpretation.

—Stephen G. Sireci

See also Criterion Validity; Face Validity; Predictive Validity; Reliability Theory; Validity Theory

Further Reading

- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3), 5–11.
- Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement*, 16, 269–282.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.
- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299–321.
- Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement*, 16, 17–31.

CONTINUOUS VARIABLE

If members of a group of people (or animals or things) are the same in terms of a particular characteristic of interest, there is no variability in the group. Usually, however, there is at least some degree of heterogeneity among the group's members. In this second, more

typical situation, the characteristic being focused on is said to be a variable.

A variable is said to be a *continuous variable* if it is theoretically possible for the group members to lie anywhere along an imaginary line segment with ends that represent small and large amounts of the characteristic. The litmus test for determining whether a particular variable is or is not a continuous variable is this pair of questions: (a) If a small difference exists between two of the people (or animals or things) in the group, is it possible for a third member of the group to be positioned between the first two? (b) If so, could the third member of the group be positioned between the first two no matter how small the difference between the first two?

To illustrate how this two-question test allows us to determine whether a variable is a continuous variable, imagine a group of people who are not all equally tall. Height, therefore, is a variable. To determine whether height is a continuous variable, first imagine two people in the group, A and B, who are nearly the same height; then, ask whether a third person, C, could be taller than A but shorter than B. Now comes the critical second question. Could C be between A and B in terms of height no matter how small the difference between A and B? The answer is “yes” so long as A and B are not the same height. Therefore, height is a continuous variable.

Other examples of continuous variables include measurements of many physical traits, such as weight, length, speed, and temperature. Many psychological traits, such as intelligence, paranoia, extroversion, and creativity, are also continuous variables. No matter how similar two rocks are in weight or how similar two people are in creativity, it is theoretically possible for a third rock to have a weight between the other two rocks or for a third person to have a level of creativity between the creativity levels of the other two people.

Continuous Versus Discrete Variables

For a variable to be a continuous variable, the characteristic being focused on must be quantitative in nature. The examples used in the previous paragraphs

are quantitative in nature and also are continuous. Many other quantitative variables, however, are not continuous. Consider these two: the number of siblings a person has or the number of working televisions inside a house. If someone has one sibling whereas someone else has two siblings, it is not logically possible for someone else to be “in between” the first two people. Likewise, if one house had three working televisions while another house had four working televisions, it is impossible for another house to have between three and four working televisions. Quantitative variables such as these are called *discrete variables*.

Data Can Make a Continuous Variable Look Like a Discrete Variable

When visiting a doctor’s office, people are typically measured in terms of their weight and their temperature. The weight data usually are whole numbers (e.g., 157 pounds) while the temperature data almost always are numbers containing one decimal place (e.g., 99.7° Fahrenheit). When such measurements are collected from groups of individuals, the data make it look as if weight and temperature are discrete rather than continuous variables.

This potential confusion as to whether a variable is continuous or discrete melts away if we think about the variable *without* focusing on the data created by attempting to measure the characteristic of interest. Thus, the fact that people’s weights are typically reported in whole numbers or that their temperatures are normally reported with numbers containing just one decimal place does not alter the fact that both weight and temperature are continuous variables. It is the characteristic being focused on—rather than any data made available by trying to measure that characteristic—that determines whether a given variable is continuous.

The Distinction Between Continuous and Discrete Variables: Does It Matter?

When statistical techniques are used to summarize or analyze data, it often makes little difference whether

the variable on which measurements have been taken is continuous or discrete. For example, the correlation between scores on two variables has the same meaning (and is arrived at using the same formula) regardless of whether the variables are continuous or discrete. Similarly, it doesn't matter whether the variable beneath one's data is continuous or discrete when a standard deviation is computed or interpreted.

Yet in certain situations, it does matter whether data are tied to a continuous variable or to a discrete variable. For example, discrete and continuous variables are treated differently in probability theory. With discrete variables, probabilities can be determined for particular *points* along the score continuum. In contrast, probabilities deal only with *intervals* along the score continuum if the variable is continuous.

—Young-Hoon Ham

See also Categorical Variable; Continuous Variable; Dependent Variable; Independent Variable

Further Reading

- Cohen, B. H., & Lea, R. B. (2004). *Essentials of statistics for the social and behavioral sciences*. Hoboken, NJ: Wiley.
- Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Allyn & Bacon.
- Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis* (5th ed.). Pacific Grove, CA: Duxbury.
- Vogt, W. P. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

CONTOUR PLOT

A contour plot (or diagram) is a two-dimensional representation of a three-dimensional surface. It consists of a set of curves called contours formed by projecting

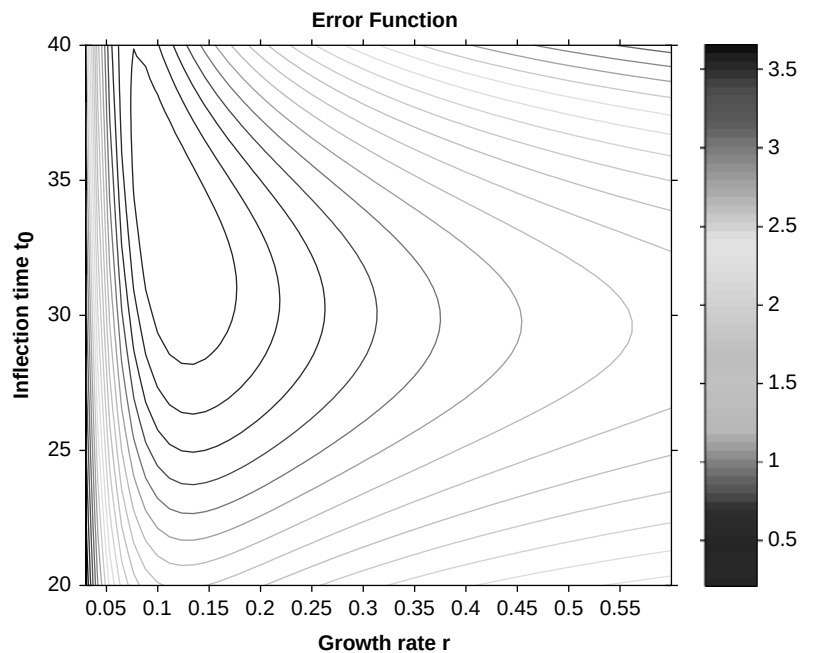


Figure 1 Contour Plot of U.S. Cumulative Underground Gas Storage Capacity

Source: Data from International Gas Consulting, Inc.

the curves of intersection of the surface with planes parallel to one of the coordinate planes. Typically, the surface is the graph of a function, which may be given as a formula or as output of measured data. The contours share the property that on any particular curve, the value of the dependent variable is constant. Normally, the diagram is such that the values between contours vary by a constant amount.

Figure 1 is an example of a contour plot. It was used to fit a logistic curve, that is, a curve of the form

$$y(t) = \frac{K}{1 + C \exp(-rt)},$$

in the least-squares sense to a set of data for the U.S. cumulative underground gas storage capacity from 1932 to 2000 (C is an arbitrary constant of integration). By manipulating the expression for a logistic function, it is possible to write a formula in terms of only two parameters: its growth rate and its inflection point. In order to minimize the corresponding least-square error function, one needs an accurate estimate of those parameters. This usually requires a

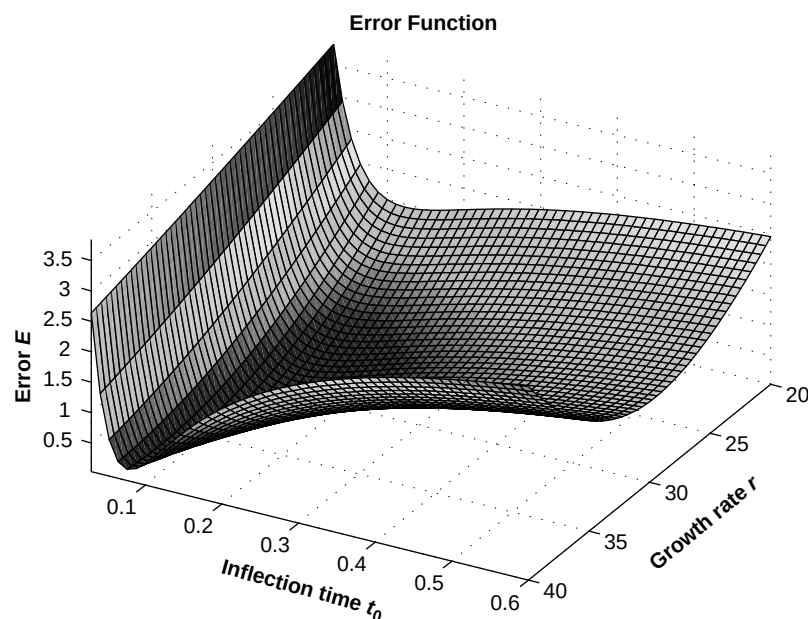


Figure 2 Surface Plot of U.S. Cumulative Underground Gas Storage Capacity

first estimate, which in this example can be obtained more easily from the contour plot of the error function than from the three-dimensional graph (see Figure 2).

As with all contour plots, the contour plot illustrated above reflects the intersection of planes parallel to the inflection time-growth rate plane (t_0, r - plane) and the surface of the error function displayed in Figure 2. Each contour depicts a curve of constant error; that is, the value of the error function remains the same along a curve.

Graphing contour lines by hand is generally impractical. Instead, one uses either graphics software or a mathematical or statistical software package. The plots shown were created in Matlab; instead of labeling values of the error at some curves, values are indicated by a scale given by a color bar.

Contour plots are used in very diverse fields; in particular, they are widely used in the earth sciences. A common application is topographic maps, which display elevation in terms of longitude and latitude. Since the elevation between any two consecutive contours varies by the same amount, proximity of contour lines indicates the rate of change of elevation.

—Silvia A. Madrid

See also Dependent Variable; Logistic Regression Analysis

Further Reading

Mooney, D., & Swift, R. (1999). *A course in mathematical modeling*. Washington, DC: MAA.

Bermuda Atlantic Time Series Study: <http://coexploration.org/bbsr/classroombats/html/visualization.html> (illustrates use of contour plots in oceanography)

Reliability Hot Wire: <http://www.weibull.com/hotwire/issue18/relbasics18.htm> (discusses contour bounds and confidence bounds on parameters)

CONVENIENCE SAMPLING

Convenience sampling is a type of survey sampling in which interviewers are allowed to choose convenient members

of the population to interview. The goal of a survey is to gather data in order to describe the characteristics of a population. A population consists of units, or elements, which, depending on the application, can be individuals, households, land areas, bank accounts, or hospital records at a specific time and location. A survey collects information on a sample, or subset, of the population. In some surveys, specific units or elements are chosen by the survey designers to be in the sample. Interviewers are assigned to interview the members of the selected sample. Sometimes multiple attempts at contacting and collecting data from the selected sample members are made. In convenience samples, on the other hand, interviewers themselves are given some latitude in selecting the population members to interview. That is, the survey designers and planners do not strictly control the selection of the sample.

Convenience samples occur in many forms. If an interviewer is told to stand on a corner or at the exit of a shopping mall and find adults to complete a survey about local schools, then this is a convenience sample because the instructions of whom to interview are not explicit. If customer satisfaction forms are distributed to certain customers in a restaurant, then this is a convenience sample if the waiters are allowed to choose

which customers to give comment forms. Internet and call-in opinion polls also could be considered convenience samples in which the sample essentially selects itself by choosing to participate. A subtler example of convenience sampling occurs when a telephone interviewer dials the telephone numbers that are randomly generated from a computer until a willing respondent is found. Although the telephone numbers are being provided to the interviewer, the interviewer is going to speak to the first available and willing people reached. This is a convenience sample because the survey planners are not picking specific numbers that should be called. In this scenario, the people called by the interviewer essentially decide themselves whether or not to be in the sample.

In the shopping mall example, it is likely that the interviewer will tend to approach people who look friendly and are walking at a casual pace rather than people who are rushing and in a visibly bad mood. Interviewers might tend to choose people of their own gender, age group, ethnicity, or race to interview slightly more often than they choose others. The interviewer also is likely to choose a desirable place to stand and to avoid loud places, such as by the door to a video arcade, and smelly locations, such as near a garbage or designated smoking area. As a result of the chosen location and tendencies to approach certain types of individuals and avoid others, it is possible that the convenience sample will not produce results that are truly representative of the entire population of interest. The convenience sample could be representative of certain subgroups in the population, but due to the lack of control by the survey planners, it is difficult to specify which population exactly is being represented.

Estimates of population characteristics based on convenience samples are affected by *selection bias*. Since the interviewers choose respondents that they want to interview or respondents decide whether or not to participate, there is a potential for selection bias. If the respondents in the survey are systematically different from the general population on the variables being measured, then estimates of characteristics will be different on average from what they would have been with a controlled probability-sampling scheme. *Probability sampling* refers to a

collection of survey sample designs in which the survey planner or researcher controls which units are in the sample and selects the sample using known probabilities of selection. The probabilities of selection can be used to produce estimates of population characteristics without the problem of selection bias. Probability sampling is the standard methodology for large-scale surveys intended to support scientific studies and decision making for government policy. That is not to say that convenience sampling should never be done. Sometimes it is very helpful to get some feedback and suggestions from people especially concerned with a problem or issue. One should be careful, however, about the limitations of general statements about a large population based on convenience samples.

—Michael D. Larsen

See also Nonprobability Sampling; Quota Sample; Random Sampling

Further Reading

- Kelly, H., Riddell, M. A., Gidding, H. F., Nolan, T., & Gilbert, G. L. (2002). A random cluster survey and a convenience sample give comparable estimates of immunity to vaccine preventable diseases in children of school age in Victoria, Australia. *Vaccine*, 20(25–26), 3130–3136.
- Schonlau, M. (2004). Will web surveys ever become part of mainstream research? *Journal of Internet Medical Research*, 6(3) article e31. Retrieved May 14, 2006, from <http://www.jmir.org/2004/3/e31/>
- Schonlau, M., Fricker, R. D., Jr., & Elliott, M. N. (2001). *Conducting research surveys via e-mail and the Web* (chapter 4). Santa Monica, CA: RAND. Retrieved May 14, 2006, from <http://www.rand.org/publications/MR/MR1480/MR1480.ch4.pdf>

Nonprobability sampling: http://www.statcan.ca/english/edu/power/ch13/non_probability/non_probability.htm

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Zelinski, E. M., Burnight, K. P., & Lane, C. J. (2001). The relationship between subjective and objective memory in the oldest old: Comparisons of findings from a representative and a convenience sample. *Journal of Aging and Health*, 13(2), 248–266.

Convenience sampling is only one of many different techniques that scientists use to select participants in a study. Elizabeth Zelinski and her colleagues tested the hypothesis that subjective memory ratings are more accurate in the oldest old than in the young old and also tested whether a representative sample was more accurate than a convenience sample. The results of analysis of subjective ratings and participant characteristics on recall were compared between a nationally representative sample of 6,446 adults ages 70 to 103 and a convenience sample of 326 adults ages 70 to 97. Researchers found that education interacted with memory ratings in the prediction of performance in the representative sample, with better prediction for more highly educated participants than for participants with lower levels of education. Neither hypothesis was supported.

COPING RESOURCES INVENTORY FOR STRESS

Coping refers to conscious or unconscious efforts to manage internal and external demands or situations that are appraised as taxing one's personal resources. Instruments measuring coping usually fall into one of two categories: those measuring coping processes and those measuring coping resources.

While coping processes are thoughts or behaviors occurring after stressful events have occurred, coping resources are factors in place before such stressors occur. Coping resources may include psychological traits, cognitive skills, belief systems, social support, physical health and fitness, and financial resources.

The Coping Resources Inventory for Stress (CRIS) is a comprehensive measure of personal resources for coping with stress. The CRIS offers 15 resource measures, an overall coping effectiveness measure, five validity keys, and a computer-generated interpretative report that suggests ways of strengthening deficit resources. The resource scales are self-disclosure, self-direction, confidence, acceptance, social support, financial freedom, physical health, physical fitness, stress monitoring,

tension control, structuring, problem solving, cognitive restructuring, functional beliefs, and social ease.

The 280 true-false items of the CRIS were distilled from more than 700 items responded to by more than 3,500 participants during a 12-year period. Contributing to the development of the inventory were disparate group studies, factor analyses, item analyses, item bias studies, reliability coefficients, and meta-analytic reviews. The normative sample ($n = 1,199$) was selected to be representative of the United States population in terms of race, gender, and age. The CRIS may be administered using test booklets or by computer.

The CRIS scales have relatively high internal consistency reliabilities (.84 to .97; $Mdn = .88$; $n = 814$), test-retest reliabilities (.76 to .95 over a 4-week period; $Mdn = .87$; $n = 34$ college students), and moderate to low intercorrelations (range .05 to .62; $Mdn = .33$). These features allow the CRIS to be used as an inventory offering stable measures of subconstructs that all contribute to one superordinate construct, coping resources. Some of the studies for establishing the validity of the inventory include measures of illness, emotional distress, personality type, drug dependency, occupational choice, acculturation, and life satisfaction. The CRIS is published by Datamax Corporation, Atlanta, Georgia.

—Kenneth B. Matheny and
William L. Curlette

Further Reading

- Matheny, K. B., Aycok, D., Curlette, W. L., & Junker, G. (1993 November). Coping Resources Inventory for Stress: A measure of perceived coping resources. *Journal of Clinical Psychology*, 49(6), 815–830.
- Matheny, K. B., & Curlette, W. L. (1998). The Coping Resources Inventory for Stress: A comprehensive measure of stress-coping resources. In C. P. Zarakett & R. J. Wood (Eds.), *Evaluating stress: A book of resources*. Lanham, MD: Scarecrow.
- Matheny, K. B., Curlette, W. L., Aysan, F., Herrington, A., Gfroerer, C. A., Thompson, D., et al. (2002). Coping resources, perceived stress, and life satisfaction among Turkish and American university students. *International Journal of Stress Management*, 9(2), 81–97.

CORRELATION COEFFICIENT

Correlation coefficient is a measure of association between two variables, and it ranges between -1 and 1 . If the two variables are in perfect linear relationship, the correlation coefficient will be either 1 or -1 . The sign depends on whether the variables are positively or negatively related. The correlation coefficient is 0 if there is no linear relationship between the variables. Two different types of correlation coefficients are in use. One is called the Pearson product-moment correlation coefficient, and the other is called the Spearman rank correlation coefficient, which is based on the rank relationship between variables. The Pearson product-moment correlation coefficient is more widely used in measuring the association between two variables. Given paired measurements $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, the Pearson product-moment correlation coefficient is a measure of association given by

$$r_p = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample mean of $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$, respectively.

Case Study and Data

The following 25 paired measurements can be found at <http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html>:

77	84
137	116
117	123
94	128
116	155
102	101
111	118
93	113
88	104

102	88
91	104
104	129
107	86
112	96
113	144
110	139
125	113
133	146
115	128
105	115
87	79
91	85
100	120
76	60
66	51

For a total of 25 occupational groups, the first variable is the smoking index (average 100), and the second variable is the lung cancer mortality index (average 100). Let us denote these paired indices as (X_i, Y_i) . The Pearson product-moment correlation coefficient is computed to be $r_p = 0.69$. Figure 1 shows the scatter plot of the smoking index versus the lung

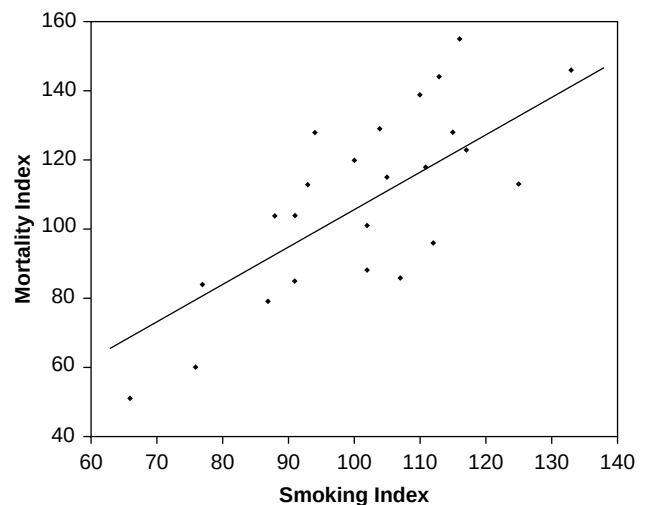


Figure 1 Scatter Plot of Smoking Index Versus Lung Cancer Mortality Index

Source: Based on data from Moore & McCabe, 1989.

Note: The straight line is the linear regression of mortality index on smoking index.

cancer mortality index. The straight line is the linear regression line given by $Y = \beta_0 + \beta_1 \times X$.

The parameters of the regression line are estimated using the least squares method, which is implemented in most statistical packages such as SAS and SPSS. The equation for the regression line is given by $Y = -2.89 + 1.09 \times X$. If (X_i, Y_i) are distributed as bivariate normal, a linear relationship exists between the regression slope and the Pearson product-moment correlation coefficient given by

$$\beta_1 \simeq \frac{\sigma_Y}{\sigma_X} r_P,$$

where σ_X and σ_Y are the sample standard deviations of the smoking index and the lung cancer mortality index, respectively ($\sigma_X = 17.2$ and $\sigma_Y = 26.11$). With the computed correlation coefficient value, we obtain

$$\beta_1 \simeq \frac{26.11}{17.20} \times 0.69 = 1.05,$$

which is close to the least squares estimation of 1.09.

Statistical Inference on Population Correlation

The Pearson product-moment correlation coefficient is the underlying population correlation ρ . In the smoking and lung cancer example above, we are interested in testing whether the correlation coefficient indicates the statistical significance of relationship between smoking and the lung cancer mortality rate. So we test. $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

Assuming the normality of the measurements, the test statistic

$$T = \frac{r_P \sqrt{n-2}}{\sqrt{1-r_P^2}}$$

follows the t distribution with $n-2$ degrees of freedom. The case study gives

$$T = \frac{0.69 \sqrt{25-2}}{\sqrt{1-0.69^2}} = 4.54.$$

This t value is compared with the 95% quantile point of the t distribution with $n-2$ degrees of

freedom, which is 1.71. Since the t value is larger than the quantile point, we reject the null hypothesis and conclude that there is correlation between the smoking index and the lung cancer mortality index at significance level $\alpha = 0.1$. Although r_P itself can be used as a test statistic to test more general hypotheses about ρ , the exact distribution of ρ is difficult to obtain. One widely used technique is to use the Fisher transform, which transforms the correlation into

$$F(r_P) = \frac{1}{2} \ln \left(\frac{1+r_P}{1-r_P} \right).$$

Then for moderately large samples, the Fisher transform is normally distributed with mean $\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$ and variance $\frac{1}{n-3}$. Then the test statistic is $Z = \sqrt{n-3} (F(r_P) - F(\rho))$, which is a standard normal distribution. For the case study example, under the null hypothesis, we have

$$\begin{aligned} Z &= \sqrt{25-3} \left(\frac{1}{2} \ln \left(\frac{1+0.69}{1-0.69} \right) - \frac{1}{2} \ln \left(\frac{1+0}{1-0} \right) \right) \\ &= 3.98. \end{aligned}$$

The Z value is compared with the 95% quantile point of the standard normal, which is 1.64. Since the Z value is larger than the quantile point, we reject the null hypothesis and conclude that there is correlation between the smoking index and the lung cancer mortality index.

—Moo K. Chung

See also Coefficients of Correlation, Alienation, and Determination; Multiple Correlation Coefficient; Part and Partial Correlation

Further Reading

- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10, 507–521.
- Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. New York: W. H. Freeman. (Original source: Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970–1972, Her Majesty's Stationery Office, London, 1978)
- Rummel, R. J. (n.d.). *Understanding correlation*. Retrieved from <http://www.mega.nu:8080/ampp/rummel/uc.htm>

Correlation coefficient page: <http://mathworld.wolfram.com/CorrelationCoefficient.html>

Spearman rank correlation coefficient: <http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html>

CORRESPONDENCE ANALYSIS

Correspondence analysis (CA) is an exploratory multivariate technique that converts data organized in a two-way table into graphical displays, with the categories of the two variables depicted as points. The objective is to construct a low-dimensional map that well summarizes the data, which in the case of CA are the associations between two categorical variables. Mathematically, the technique decomposes the χ^2 measure of association of the two-way table into components in a manner similar to that of principal component analysis for continuous data. In CA, no assumptions about the data-generating mechanism are made, a significant departure from log linear analysis. The primary objective of the technique is the representation of the underlying structure of the observed data.

CA can be traced back to the work of Hirschfeld (see de Leeuw). It has been rediscovered in various forms by Fisher, Guttman, Hayashi, and Benzécri, who emphasized the geometric aspects of the technique. Extensive expositions and a discussion of its similarities to and differences from other methods, such as dual scaling and canonical correlation, can be found in the books by Nishisato and Greenacre.

There are two variants of the technique: *simple* CA, which deals with two-way tables, and *multiple* CA, a generalization designed to handle more than two categorical variables.

Simple CA

Consider two categories variables V_1 and V_2 , with I and J categories, respectively. Let α_2 denote the corresponding $I \times J$ two-way table, whose entries $x_{i,j}$ contain counts of the co-occurrences of categories i and j , and let N denote the

grand total of $\mathbf{X}$ (i.e., $N = \sum_{i,j} x_{i,j}$). Let $\mathbf{Z}$ denote the probability matrix obtained as $N^{-1}\mathbf{X}$. Further, let $\mathbf{r} = \mathbf{Z}\mathbf{1}$ denote the vector of row marginals and $\mathbf{c} = \mathbf{Z}^T\mathbf{1}$ the vector of column marginals. Finally, define $\mathbf{D}_r = \text{diag}\{\mathbf{r}\}$ as the diagonal matrix containing the elements of vector $\mathbf{r}$, and similarly define $\mathbf{D}_c = \text{diag}\{\mathbf{c}\}$.

An example that illustrates the notation of a two-way table first used by Fisher is shown in Table 1. Data on 5,387 school children from Caithness, Scotland, were collected for two categorical variables, eye color and hair color.

The dependencies between the rows (columns) of $\mathbf{Z}$ can be captured by the so-called χ^2 distances defined (here between row i and i') as

$$d_{\chi}^2(i, i') = N \sum_{j=1}^J \frac{1}{c_j} \left(\frac{z_{i,j}}{r_i} - \frac{z_{i',j}}{r_{i'}} \right)^2. \quad (1)$$

Equation 1 shows that the χ^2 distance is a measure of the difference between the profiles of rows i and i' . It also shows that by correcting the entries in the table by the row marginals, proportional row profiles yield zero distances. Also note that squared differences between row categories i and i' are weighted heavily if the corresponding column marginal is small, while such differences contribute little to the distance measure if the column marginal is large.

The objective of CA is to approximate the χ^2 distances by Euclidean distances in some low-dimensional space. In order to derive the $I \times L$ coordinates $\mathbf{F}$ (with $L = 2$ or 3 for visualization purposes) in the new Euclidean space, we consider the singular value decomposition of the observed frequencies minus the expected frequencies, corrected for row and column marginals:

Table 1 Example of Two-Way Table

		$V_2 = \text{Hair color}$				
$V_1 = \text{eye color}$		Fair	Red	Medium	Dark	Black
	Total (r)					
Light	688	116	584	188	4	1,580
Blue	326	38	241	110	3	718
Medium	343	84	909	412	26	1,774
Dark	98	48	403	681	85	1,315
Total (c)	1,455	286	2,137	1,391	118	$N = 5,387$

$$\mathbf{D}_r^{-\frac{1}{2}} (\mathbf{Z} - \mathbf{r}\mathbf{c}^T) \mathbf{D}_c^{-\frac{1}{2}} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T. \quad (2)$$

The optimal coordinates of the row categories are given (after normalization) by

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{P}, \quad (3)$$

so that $\mathbf{F}^T \mathbf{D}_r \mathbf{F} = \mathbf{I}$ and $\mathbf{1}^T \mathbf{D}_r \mathbf{F} = \mathbf{0}$ (i.e., in each dimension the row scores have a weighted variance of 1 and a weighted average of 0; note that several implementations of CA normalize the row factor scores such that their variance is equal to the corresponding eigenvalue; in this case the factor scores are computed as $\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{P}\mathbf{\Delta}$).

Since Euclidean distances between the $\mathbf{F}$ coordinates of the row categories approximate the original χ^2 distances, it can be seen that when two categories are depicted close together, their profiles are similar, while they are different if their positions are far apart. Finally, when a row category is near the center of the $\mathbf{F}$ space, its profile is similar to that of the corresponding column marginal.

Given the above configuration for the row categories, we can compute the column category configuration as follows:

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{Q}\mathbf{\Delta}, \quad (4)$$

which implies that the column points are in the center of gravity of the row points. Since the analysis is symmetric regarding rows and columns, one could have calculated χ^2 distances between the column categories and obtained the corresponding representation of row and column categories in L -dimensional Euclidean space. Some algebra shows that the Pearson's χ^2 statistic used for testing independence between two categorical variables is related to the formulas for CA by

$$\text{trace} \{ \mathbf{\Delta}^2 \} = \frac{X^2}{N}, \quad (5)$$

known as the *total inertia* in the literature.

An illustration of CA using Fisher's data from Table 1 is shown in Figure 1.

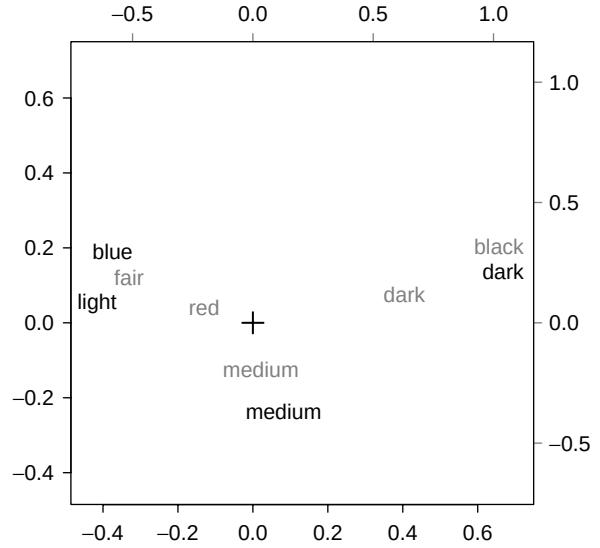


Figure 1 Correspondence Analysis Using Fisher's Data

It can be seen that there is high association between fair hair and light and blue eye color and similarly between black hair and dark eyes. On the other hand, dark hair is associated to a large extent with both dark and medium-colored eyes. Further, students with medium hair color have mostly medium eye color but can also have all three other eye colors.

Multiple Correspondence Analysis

In the presence of more than two categorical variables, there are two possible ways to proceed. We examine both of them next. Suppose that data have been collected on I objects and K categorical variables, with J_k categories per variable. Let $\mathbf{X}_k$ be a $I \times J_k$ indicator (binary) matrix with entries $x_{i,j} = 1$ if object i belongs to category j of variable k and 0 if not. Let $\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2 | \dots | \mathbf{X}_K]$ be the *superindicator* matrix obtained by concatenating the indicator matrices of all K variables. The symmetric matrix $\mathbf{B} = \mathbf{X}^T \mathbf{X}$, known as the *Burt table*, contains the marginals for the categories of all variables along the main diagonal, together with all two-tables of the K variables in the off-diagonal.

Applying simple CA to the Burt table yields the following solution for the category points:

$$\mathbf{D}_B^{-\frac{1}{2}} (\mathbf{B} - \mathbf{M}^{-1} \mathbf{D}_B \mathbf{1} \mathbf{1}^T \mathbf{D}_B) \mathbf{D}_B^{-\frac{1}{2}} = \mathbf{P}\mathbf{A}\mathbf{P}^T, \quad (6)$$

where $\mathbf{D}_B = \text{diag}\{\mathbf{B}\}$ and $N = I \times K^2$ (i.e., N is the grand total of $\mathbf{B}$). The coordinates of the category points are given by

$$\mathbf{F} = \sqrt{N} \mathbf{D}_B^{-\frac{1}{2}} \mathbf{P} \mathbf{A}. \quad (7)$$

Multiple CA is the joint analysis of all the two-way tables of K categorical variables.

In many cases, one may want a joint representation of both objects and categories of variables they belong to. This objective can be achieved by applying simple correspondence analysis directly to the superindicator matrix $\mathbf{X}$. The coordinates $\mathbf{F}$ for the objects are given by the singular value decomposition of

$$(\mathbf{I} - \mathbf{I}^{-1} \mathbf{1} \mathbf{1}^T) \mathbf{X} \mathbf{D}_B^{-\frac{1}{2}} = \mathbf{P} \mathbf{A} \mathbf{Q}^T, \quad (8)$$

where the left-hand side corresponds to the superindicator matrix of the data expressed in deviations from the column means and weighted by the variables' marginal frequencies. Specifically, $\mathbf{F} = \sqrt{\mathbf{I}} \mathbf{P}$ (i.e., the coordinates have unit variance). The coordinates of the categories can be obtained by $\mathbf{G} = \mathbf{K} \mathbf{D}_B^{-1} \mathbf{X}^T \mathbf{F}$, which shows that a category point is located in the center of gravity of the objects that belong to it. An alternative derivation of the above solution as a graph-drawing technique is given in Michailidis and de Leeuw, and various extensions are discussed in Van Rijkevorsel and de Leeuw.

Multiple CA is illustrated in Figure 2 on a small data set of 21 objects (sleeping bags) and three categorical variables (price, comprised of three categories; fiber type, with two categories; and quality, with three categories).

From the joint map of objects and categories (connected by lines), it becomes apparent that there are good, expensive sleeping bags filled with down fibers and cheap, bad-quality ones filled with synthetic

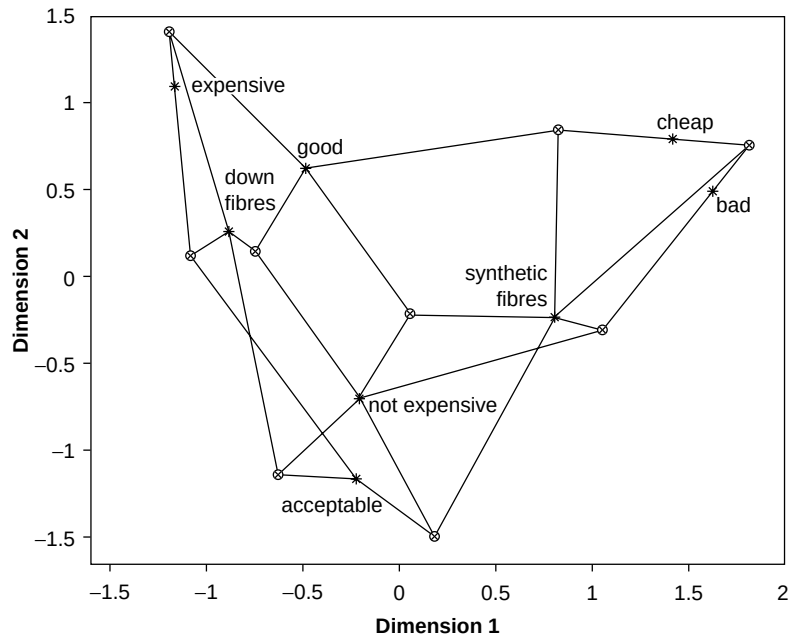


Figure 2 Multiple Correspondence Analysis

fibers. Further, there are some not particularly expensive sleeping bags of acceptable quality made of synthetic or down fibers.

—George Michailidis

See also Discriminant Analysis; Discriminant Correspondence Analysis; Distance; DISTATIS; Factor Analysis; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Multiple Factor Analysis; STATIS

Further Reading

- Benzécri, J. P. (1973). *L'analyse des données*. Paris: Dunod.
- De Leeuw, J. (1983). On the prehistory of correspondence analysis. *Statistica Neerlandica*, 37, 161–164.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, UK: Wiley.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 5, 121–143.
- Michailidis, G., & de Leeuw, J. (2000). Multilevel homogeneity analysis with differential weighting. *Computational Statistics and Data Analysis*, 32, 411–442.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: Toronto University Press.

Van Rijckevorsel, J., and de Leeuw, J. (Eds.). (1988). *Component and correspondence analysis*. Chichester, UK: Wiley.

COVARIANCE

Covariance is a measure of covariation, or linear relationship, between two variables that is expressed in the original units of measurement. Equation 1 formalizes this definition for the population covariance. Examination of the equation shows that covariance is a bivariate statistic that quantifies the joint dispersion of two variables from their respective means. Equation 1 also shows that covariance is the average of two sets of deviation scores:

$$\begin{aligned} E[(X - \mu_x)(Y - \mu_y)] \\ = \frac{\sum (X - \mu_x)(Y - \mu_y)}{N} = \sigma_{xy}. \end{aligned} \quad (1)$$

The sample covariance is a statistic that estimates the degree to which two variables covary in the population. The sample covariance is an unbiased estimate of population covariance when observational pairs are sampled independently from the population. The sample covariance presumes that the functional relationship between the two variables is linear.

Equation 2 is the deviation score formula for the sample covariance for two variables X and Y :

$$COV_{XY} = s_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}. \quad (2)$$

Steps to computing an unbiased estimate of the population covariance include (a) calculating the mean of the X scores and the mean of the Y scores, (b) calculating the deviation of the X and Y scores from their respective means, (c) calculating the product of the deviations for each pair of values (i.e., calculating the cross products), (d) summing these cross products, and (e) dividing the sum of the cross products by

$n - 1$ (degrees of freedom). Equation 3 shows that the covariance of a variable with itself is the variance:

$$\begin{aligned} s_{xx} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{n - 1} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = s^2. \end{aligned} \quad (3)$$

The units of measurement of the covariance are not intuitive, because they are expressed in terms of the cross products of the scales of the X and Y scores. It is difficult to interpret covariance because it is very difficult to think in terms of a statistical value as summarizing how two different and frequently arbitrary metrics covary (e.g., IQ points and Graduate Record Examinations points).

Properties of the Covariance

Covariance has no limits and can take on any value between plus and minus infinity. Negative values indicate that high scores on one variable tend to be associated with low scores on the other variable. Positive values indicate that high scores on one variable tend to be associated with high scores on the other variable. A covariance of 0 indicates that there is no linear relationship between the two variables. Because covariance measures only linear dependence, covariance values of 0 do not indicate independence.

The sample covariances, along with sample sizes, sample means, and sample variances, form the building blocks of the general linear statistical model. Essentially, the covariance summarizes all the important information contained in a set of parameters of a linear model. Covariance becomes more positive for each pair of values that differ from their mean in the same direction, and it becomes more negative with each pair of values that differ from their mean in opposite directions. The more often the scores differ in the same direction, the more positive the covariance, and the more often they differ in opposite directions, the more negative the covariance.

Equation 4 shows that correlation is merely a scale-free measure of covariance or linear relationship; that is, correlation is merely a standardized covariance. When covariance is calculated, the sample means are subtracted from the scores, and when we calculate the correlation, we divide the covariance by the product of the standard deviations. This shows that the correlation is the covariance of the z_x and z_y scores,

$$r_{xy} = \frac{\sum_{i=1}^n z_x z_y}{N}.$$

Clearly, the essence of correlation, that it measures both the strength and the direction of the linear relationship, is contained in the covariance:

$$\begin{aligned} \text{correlation} = r_{xy} &= \frac{COV_{XY}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / n - 1}{s_x s_y}. \end{aligned} \quad (4)$$

—Ward Rodriguez

See also Analysis of Covariance (ANCOVA); Linear Regression; Regression Analysis; Variance

Further Reading

- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Stilson, D. W. (1966). *Probability and statistics in psychological research and theory* (pp. 251–255). San Francisco: Holden-Day.
- Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate* (pp. 36–38). Upper Saddle River, NJ: Prentice Hall.

CRITERION-REFERENCED TESTS

Although criterion-referenced tests (CRTs) can be developed to measure performance at the domain level (e.g., mathematics, reading), they are much more commonly used to measure mastery of short-term objectives (e.g., a unit on the Civil War). As classroom

tests, CRTs have existed for many decades, if not centuries. The work of Mager and Glaser in the early 1960s started an important and continuing movement to improve the way that educators estimate students' achievement. R. F. Mager wrote a popular book that motivated educators to be precise in explicating the skills they wanted their students to learn. Robert Glaser, who is generally credited with coining the term *criterion-referenced test*, initiated the movement to measure the mastery of instructional objectives with reliability and validity.

What Is a CRT?

A CRT is a measure designed to estimate mastery of an identified unit of a curriculum (e.g., battles of the Civil War, multidigit addition with regrouping, use of prepositions). CRTs are also referred to as *curriculum-based measures* and more broadly as *curriculum-based assessment*. CRTs are standardized instruments, which are constructed with sufficient precision that different examiners will administer, score, and interpret results in the same way. CRTs contain items designed to represent the unit of instruction adequately. Each item has a predetermined correct answer that can be scored objectively by the assessor. A CRT is used for two main purposes. First, it is used to determine whether a student is weak in a given skill and needs further instruction. Second, it is used to determine the effectiveness of instruction. Although CRTs are seldom normed nationally, it is beneficial to collect local norms for appropriate grade groups.

In contrast to norm-referenced tests, which use relative mastery criteria to interpret scores, CRTs use absolute mastery criteria. Therefore, the student's performance is not compared to that of other students but to a predetermined absolute standard of performance. Most commonly, CRTs measure performance as percentage correct. The particular measure used should be based on real-life demands. Because CRTs are usually used to measure short-term objectives, they tend to be formative rather than summative in nature. Thus, for skills at the lowest taxonomic levels (e.g., miniskills), educators may obtain mastery estimates on a weekly or even a daily basis.

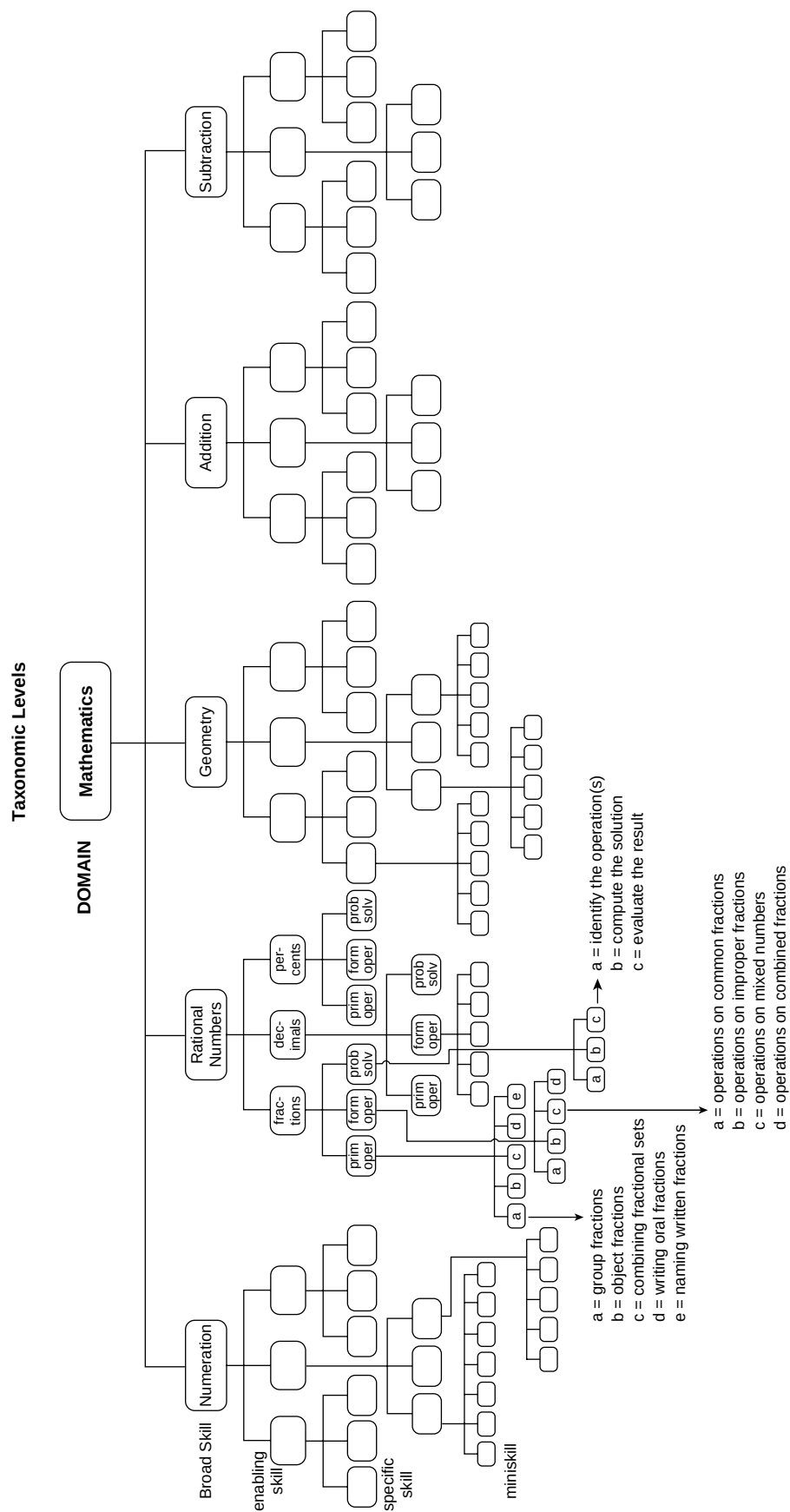


Figure 1 A Partial Taxonomy for Mathematics

Some writers distinguish between domain-referenced tests and objective-referenced tests. Domain-referenced tests provide a measure that is interpretable in the context of a clearly defined and delimited set of objectives that collectively constitute an entire domain (e.g., mathematics, reading). Objective-referenced tests provide a measure that is interpretable only for a particular instructional objective.

Selection of Taxonomies of Objectives

It is useful for practitioners to select taxonomies of objectives for domains relevant to their practice. For instance, a mathematics teacher would benefit from a complete specification of the broad, enabling, and specific math skills needed by American citizens. Such a taxonomy represents the results of a task analysis of a domain. Although there is no such thing as a perfect taxonomy, national organizations (e.g., the National Council of Teachers of Mathematics, National Council of Teachers of English, National Research Council-Science) often publish taxonomies. A layout of a partial math taxonomy might look like that displayed in Figure 1.

The layout should be accompanied by a document containing instructional objectives for each skill represented in the layout.

Here is a set of objectives for basic addition facts:

3.1.0 Given all possible permutations of the addition facts (addends 0–9) presented vertically and in random order, the child will write answers with at least 95% accuracy and at a rate of at least 20 correct per minute.

3.1.1 Given all possible permutations of the addition facts (addends 0–5) presented vertically and in random order, the child will write answers with at least 95% accuracy and at a rate of at least 25 correct per minute.

3.1.2 Given all possible permutations of the addition facts (addends 6–9) presented vertically and in random order, the child will write answers with at

least 95% accuracy and at a rate of at least 20 correct per minute.

Objective 3.1.0 represents an enabling skill, enabling in the sense that mastery will enable the student to address higher-order skills in addition and in the domain of mathematics.

Notice that the objective contains the three critical features of a good objective:

1. A statement of the conditions under which the student will perform (i.e., “given all possible permutations of the addition facts [addends 0–9] presented vertically and in random order”)
2. The behavior (i.e., “the child will write answers”)
3. Criteria for minimally acceptable performance (i.e., “with at least 95% accuracy and at a rate of at least 20 correct per minute”; in this case, criteria for both accuracy and rate are specified)

When these three features are clearly explicated in the objective, it is easier for the test writer to construct items. For objective 3.1.0, there should be 100 items, exhausting every possible permutation of the addends 0 through 9. They should be randomly sequenced, and they should be presented vertically, not horizontally (e.g., $5 + 7 = \underline{\quad}$). There can be little question of the content validity of the test because all 100 single-digit addition facts are represented, though some may argue that horizontally presented items should be included. The student will write answers, so it will be a paper-and-pencil test. Finally, the student must complete the test within 5 minutes and miss no more than five items.

Construction of CRTs

The construction of a CRT is a multistage process. Although it requires time and effort, the process is well worth the dividends it pays in accurate and valid decision making. The alternative, which is of little value, is to create hurriedly a weak instrument, which estimates mastery poorly and should never be used at all. A well-constructed CRT is worth filing for future use, though we recommend the creation of several alternate forms of tests that cover the same curriculum unit. Below is a presentation of each step in the development of a CRT.

Naming the Test

Although it seems trivial, it is important to give the test a name that accurately represents its content. Over the course of years, teachers construct many tests, and a filing system that allows efficient retrieval for future use demands that test names obviously reflect their content. For instance, the name *phonology* implies a lot: sounding out vowels, consonants, consonant blends, digraphs, and diphthongs in the initial, medial, and end positions in a word. In other words, the name implies that everything subsumed under the broad skill of phonology is measured by the test. If the test covers only consonant blends in the medial position, the name *phonology* misrepresents the test's content.

Objective(s) Represented by the Test

It is important that tests be created *after* objectives have been specified. The test items are then constructed to reflect mastery of the objectives, not the other way around. Ideally, the objectives will be drawn from a large taxonomy of objectives that covers the entire domain.

Conditions under which the behavior will occur include any condition that might influence the examinee's performance during testing. Many conditions are generic and cover nearly all testing situations: good lighting, comfortable temperature, seats and tables of the correct size, and so on. Other conditions are unique to a particular test: necessary materials, explicit instructions to the examinee, timing factors, and others. The statement of the behavior should ideally include an action verb (e.g., write, pronounce orally, walk). In other words, the behavior should be objectively defined such that two people would agree that the behavior has or has not occurred. Criteria should be in the form of one of a number of recognized scores: percentage correct, behavior rates, duration, response latency, intensity, standard score, percentile rank, and so on.

Instructions for Administration

This component tells the user how the test should be given. It elaborates the conditions element of the

objective. The purpose of instructions for administration is to standardize data collection so that from occasion to occasion, from child to child, and from examiner to examiner, the test is administered in the same way. This makes the scores comparable. Typical elements included are (a) instructions to the child, (b) materials needed (e.g., two sharpened number 2 pencils, a watch for timing purposes), (c) ways to deal with interruptions, and (d) ways to deal with questions from the child. The test makers must ask themselves what elements impinge on the successful administration of the test.

Instructions for Scoring

This section tells the user how to transform the examinee's responses into item scores and total scores. This often means providing criteria for correct and incorrect responses to individual items in a scoring key. A formula may be required to obtain a total score (e.g., the formula for behavior rates), and it should be illustrated for the uninformed user.

Instructions for Interpretation

Here the user is told how to make decisions on the basis of the score(s) obtained from an administration of the instrument. Basically, the criteria for minimally acceptable performance laid out in the objective guide this process. If the criterion mentions 95% accuracy, then the user should compare the examinee's score with 95%. If the examinee's score equals or exceeds that value, the child has mastered the objective. If not, then the objective needs more instruction and practice.

Specific Items in the Instrument

The key here is for the test maker to ensure that the items in the test are representative of the skills specified in the objectives. First, there must be enough items to comprise a reliable sample of the skills in question. It is rarely possible to have a reliable measure of any objective with less than 25 items. Second, the items should adequately represent

Table 1 Estimated Standard Errors of Test Scores

Number of items	Standard error ^a	Exceptions: Regardless of the length of the test, the standard error is
< 24	2	0 when the score is zero or perfect
24–47	3	1 when 1 or 2 percentage points from 0 or 100%
48–89	4	2 when 3 to 7 percentage points from 0 or 100%
90–109	5	3 when 8 to 15 percentage points from 0 or 100%
110–129	6	
130–150	7	

Source: Used with permission ©1979 Ronald C. Eaves.

a. Standard errors are in raw score form. Items are assumed to be scored dichotomously (i.e., 0 or 1).

the various kinds of subskills contained within an objective. For instance, a test on addition facts is unrepresentative if it does not include items containing 6 or 8.

Standard Error of Measurement

The standard error of measurement (SEM) of a test can be estimated using the number of items contained in the test. This estimate should be included along with the other components of the test. As an example of the use of the SEM, consider the student who obtains a raw score of 7 on a test containing 11 items. The student's percentage correct is 64%. Because the percentage does not fall into one of the exceptions, the estimated SEM is 2 (for tests with less than 24 items). In order to construct a 95% confidence interval, the assessor should double the SEM (i.e., $2 \times 2 = 4$). Next, the product is subtracted from the student's raw score ($7 - 4 = 3$), and then the product is added to the student's raw score ($7 + 4 = 11$). These values represent the 95% confidence interval in raw-score form (i.e., 3–11). In percentage-correct form, the assessors can say, with the knowledge that they will be correct on 95 out of 100 such judgments, that the student's true score is contained within the interval of 27%–100%. Notice that such results on a test with few items provide virtually no useful information for decision making. The same relative performance on a

100-item test would result in a 95% confidence interval of 54%–74%.

—Ronald C. Eaves and Suzanne Woods-Groves

See also Computerized Adaptive Testing; Standards for Educational and Psychological Testing

Further Reading

- Eaves, R. C., McLaughlin, P. J., & Foster, G. G. (1979). Some simple statistical procedures for classroom use. *Diagnostique*, 4, 3–12.
- Glaser, R., & Klaus, D. J. (1962). Proficiency measurements: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in systems development*. New York: Holt, Rinehart & Winston.
- Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Macmillan.
- Mager, R. F. (1984). *Preparing instructional objectives* (2nd ed.). Belmont, CA: David S. Lake.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.
- National Council of Teachers of Mathematics standards for mathematics: <http://standards.nctm.org/index.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Laugksch, R. C., & Spargo, P. E. (1996). Construction of a paper-and-pencil Test of Basic Scientific Literacy based on selected literacy goals recommended by the American Association for the Advancement of Science. *Public Understanding of Science*, 5(4), 331–359.

This study describes the construction and validation of a **criterion-referenced** paper-and-pencil Test of Basic Scientific Literacy (TBSL) specifically designed for high school students entering university-level education in South Africa. The scientific literacy test items, designed to be answered *true*, *false*, or *do not know*, are based on a pool of 472 items developed previously from selected literacy goals recommended by the AAAS in *Science for All Americans*. Test items were pilot tested on 625 students and were included in the 110-item TBSL on the basis of item discrimination, item difficulty,

and student feedback. The TBSL consists of subtests on three constitutive dimensions of scientific literacy: the nature of science, science content knowledge, and the impact of science and technology on society. About 260 members of various South African professional science and engineering associations participated in setting a performance standard for each of the three dimensions of scientific literacy. The internal consistency of the individual TBSL subtests, and the reliability of mastery-nonmastery classification decisions based on the performance standard, was found to be about 0.80. The reliability of the overall 110-item TBSL was 0.95.

CRITERION VALIDITY

Criterion-related validity refers to the extent to which one measure estimates or predicts the values of another measure or quality. The first measure is often called the estimator or the predictor variable. The second measure is called the criterion variable in cases when a decision must be made or when the measure is regarded as valid. In some cases, neither measure has well-developed validity evidence, and there is no genuine criterion variable. There are two types of criterion-related validity: concurrent validity and predictive validity. The simple distinction between the two types concerns the time interval between obtaining the first and the second set of measurements. For concurrent validity, the data from both measures are collected at about the same time. For predictive validity, the data from the criterion measure are collected some period of time after the data of the predictor variable.

Concurrent Validity

When a developer designs an instrument intended to measure any particular construct (say, intelligence), one of the most straightforward ways to begin establishing its validity is to conduct a concurrent validity study. The basic task is to identify an available instrument (say, the Stanford-Binet Intelligence Scale), the

validity of which has already been established, that measures the same construct or set of constructs as the new instrument. The second job is for the developer to identify a large, ideally random sample of people who are appropriate for the purposes of the instrument. Next, data are collected for both instruments from every person in the sample, generally on the same day. In many circumstances, it is considered important to collect the data in a counterbalanced fashion in order to control for practice effects. Fourth, a correlational technique appropriate to the scale of measurement is applied to the pairs of scores (e.g., Pearson product-moment correlation, Spearman rank correlation). Finally, the developer compares the obtained correlation or correlations with those of the established instrument and other similar instruments collected under similar circumstances. To the extent that the results compare favorably with other available validity coefficients (as they are called), the developer has begun the lengthy process of establishing the validity of the new instrument.

In the case of norm-referenced instruments, a second analysis of the data is in order: the comparison of the mean scores for the two instruments. Although the validity coefficient, if sufficiently high, establishes the fact that the instruments measured similar constructs, at least in the current sample, it does not indicate that the new instrument is unbiased. If the standard score means are significantly different statistically, then the two instruments will lead to different decisions in practice. For instance, if the instruments both measure intelligence, and the IQ cutoff score for mental retardation is 70, a substantial number of individuals will be misdiagnosed by the new instrument. If the mean IQ is too high compared with the established intelligence test, then a considerable number of Type II errors will be made. If the opposite circumstance exists, then Type I errors are the problem. Commonly, the source of this problem is attributed to the unrepresentativeness of the new instrument's norm group.

Three additional circumstances exist in which concurrent validity studies are conducted. When an existing instrument is substantially revised, it is important for the developer to show that the new instrument

measures the relevant construct in the same way as the old instrument. This was perhaps the most important question users asked when the Vineland Adaptive Behavior Scales replaced the Vineland Social Maturity Scale. When a new instrument attempts to measure a construct that has not previously been measured with a high degree of validity (i.e., there is no available valid criterion measure), the developer may embark on a series of concurrent validity studies to determine just what the new instrument actually measures. For example, a general measure of arousal might be compared with measures of attention, persistence, curiosity, preference for novelty, vigilance, and pupillary dilation. Finally, the developer of an instrument intended to predict some future performance may be unable or unwilling to wait for some lengthy period in order to obtain an appropriate criterion measure.

Predictive Validity

Predictive validity *must* be demonstrated when an instrument is specifically designed to estimate future performance. Although this need is obvious when applicants are selected for jobs or college enrollment, many instruments in fact serve a predictive function, particularly when the construct measured is known to be stable (e.g., intelligence, academic achievement, physical strength).

Consider a newly developed test (say, the Primary School Readiness Test), the purpose of which is to predict which kindergarten children will succeed and which will fail if allowed to move on to first grade. The primary and perhaps the only purpose of such a test is to make a simple prediction: who should be retained in kindergarten and who should move on to first grade. Whatever other evidence is available, the absence of predictive validity evidence for the Primary School Readiness Test would cause any potential users to look elsewhere for their instrument of choice.

The fundamental research design for predictive validity is the same as that described above for concurrent validity, but with one difference: a time interval between the collection of the predictor variable

data and the collection of the criterion variable data. The length of the interval depends on the prediction to be made. For instance, the Primary School Readiness Test claims to predict success and failure in achieving first-grade curriculum objectives, which can best be determined at the end of first grade. Thus, a 1-year interval is called for. For tests designed to select high school students for college enrollment (e.g., the Scholastic Aptitude Test), the criterion measure is often final college grade point average. In this case, the time interval is 4 years or more. Because as time passes, performance levels can change, concurrent validity coefficients will almost always be higher than predictive validity coefficients, and shorter intervals will lead to higher predictive validity coefficients than longer intervals will.

An alternative to the conventional predictive validity study is the postdictive design. The test developer, who may be reluctant to wait a year or more to obtain predictive validity data, can sometimes use extant criterion data that were collected at some time in the past. For instance, to show that an intelligence test can effectively estimate academic achievement, the developer may correlate achievement data collected the previous year with currently obtained data from the intelligence test. Although the usual sequence of the data collection is reversed, Jensen argued that the resulting validity coefficients provide unbiased estimates of the criterion-related validity of the predictor.

Classification Accuracy

Classification accuracy studies provide another way of estimating the criterion-related validity of an instrument. Consider a test (the Always Accurate Autism Test, or AAAT) that claims to classify children with autistic disorder. If such a test were administered to 50 children previously classified as autistic (i.e., true positives) and 50 children previously classified as not autistic (i.e., true negatives), the test would ideally classify all 100 children accurately. That is seldom the case. Sensitivity is a term that refers to the percentage of true positives identified by the test, and specificity refers to the percentage of true negatives identified by the test. An inspection of Table 1 shows that the

Table 1 Classification Accuracy of the Always Accurate Autism Test

AAAT	Clinical Classification	
	Autistic (<i>n</i> = 50)	Not Autistic (<i>n</i> = 50)
Autistic	30 (60%)	22 (44%)
Not autistic	20 (40%)	28 (56%)
Overall classification accuracy	58 (58%)	

AAAT correctly identified 60% of the children with autism (i.e., sensitivity), and 56% of the children who did not have autism (specificity), and had an overall classification accuracy of 58%. Consequently, the classification accuracy of the AAAT was only a bit better than chance assignment.

The Problem With Criterion Measures

In many cases, the estimator or predictor variable (e.g., a newly developed test) may have good validity but suffer because the available criterion measures are not very valid. Although a record of performance (e.g., number of widgets made, rate of typing errors, number of tires changed) is appealing for many jobs, for many other situations (e.g., teachers, physicians, administrative assistants), no simple measure of production or performance is available. The criterion measures that are available (e.g., supervisor ratings) may be unreliable and consequently limited in their validity. For example, college grade point averages lack validity for a variety of reasons (e.g., restriction of range of college-student ability levels). This state of affairs guarantees a reduced validity coefficient between the estimator or predictor variable and the criterion variable even though the estimator or predictor variable is highly valid. In such instances, a commonly used technique is correction for attenuation. This procedure adjusts the reliability of the criterion measure so that it contains no error, providing a boost for the resulting validity coefficient and

answering the question, What would the validity coefficient be if the criterion measure were perfectly reliable?

—Ronald C. Eaves and Suzanne Woods-Groves

See also Attenuation, Correction for; Predictive Validity; Validity Coefficient; Validity Theory

Further Reading

- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Ice, G. H., & Yogo, J. (2005). Measuring stress among Luo elders: Development of the Luo Perceived Stress Scale. *Field Methods*, 17(4), 394–411.

This study describes the development of the Luo Perceived Stress Scale (LPSS) tested on 200 Luo elders. The LPSS consists of 23 emotions and uses alternating “local idioms of distress” and well-being. Due to the low level of education of the population, a yes-no format is used instead of a Likert-type scale. The scale was tested among 200 Luo elders and was found to be internally reliable. **Criterion validity** was examined through the associations between LPSS score and caregiving, social networks, depression, and cortisol. Known group validity was examined through comparisons of caregiving groups, genders, marital status, and participation in social groups. While these variables were generally associated with LPSS in the predicted direction, subsequent factor analysis suggested that the LPSS did not represent a single domain.

CRITICAL VALUE

The critical value is the value needed for rejection of the null hypothesis when it is true. For example, if the critical value for a particular type of statistic’s distribution and the sample size at a particular level

Table 1 Critical Values

Degrees of freedom	Critical value for rejection of the null hypothesis at the .05 level of significance	Critical value for rejection of the null hypothesis at the .01 level of significance
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617

of Type I error is 3.45, then the observed or computed value must exceed 3.455 for the outcome to be significant.

The following points about critical values should be remembered:

1. There is a different critical value depending on the size of the sample being evaluated (reflected in the degrees of freedom) and the Type I error set by the experimenter.
2. Critical values are usually shown in tables that accompany many statistical texts. A typical table, such as the one used for testing the difference between two independent means, would appear like Table 1. Note that the research hypothesis can be tested at both the .01 and .05 levels of significance.
3. Each test statistic (such as the *t* test or the *F* test) has its own distribution of critical values.
4. With the advent of computerized statistical analysis programs, critical values are no longer needed for comparison's sake. Rather, the exact probability of an outcome (the Type I error level) is printed out. For example, instead of the statement "The results were significant beyond the .05 level," a more accurate statement might be, "The probability of a Type I error was .043."

—Neil J. Salkind

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

CRONBACH, LEE J. (1916–2001)

Lee J. Cronbach, known widely as the creator of Cronbach's alpha, was a man whose deep interests in psychological testing and educational psychology combined as the focus of more than 50 years of research into measurement theory, program evaluation, and instruction.

Cronbach was born in Fresno, California, and came to the attention of Blanche Cummings, a follower of Lewis Terman, when his precociousness at the age of 4 made itself known as she overheard him calculating the price of potatoes at a grocery store. Ms. Cummings gave the Stanford-Binet to Cronbach in 1921 and ensured wide publicity of his score of 200. According to Cronbach, this number was inflated; however, his future eminence was not, as evidenced by his many lifetime contributions.

Because of his mother and Ms. Cummings, Cronbach's education began, not with kindergarten, but with the second grade. He graduated from high school at 14 and immediately went to college, graduating at age 18. Interestingly, Cronbach's first higher educational interests lay in chemistry and mathematics; had not a lack of funds kept him at Fresno State, he may never have discovered his passion for educational research. He went on to Berkeley for a master's in education, gaining his teaching credentials at the same time. Cronbach then taught high school (math and chemistry) while he finished an education doctorate at the University of Chicago, graduating in 1940. He married Helen Claresta Bower while at Chicago, and their five children arrived between 1941 and 1956.

While at the University of Chicago, he became a research assistant for Ralph Tyler in the Eight-Year Study, which looked into how high school curriculum affected students' success in both admission to and graduation from colleges and universities. On graduation, Cronbach worked as an associate professor at Washington State University, where he taught myriad psychology courses. Toward the end of World War II, he worked as a military psychologist at the Naval

Sonar School in San Diego, California. Cronbach went back to the University of Chicago as an associate professor and then on to the University of Illinois as a full professor. In 1964, he landed in Palo Alto, California, where he taught, conducted research, and finally retired at Stanford University, gaining the prestigious Vida Jacks Professor of Education Emeritus honor (among many others) during his tenure.

Cronbach's most famous contributions to the area of tests and measurements were his efforts at strengthening tests and measures through a deeper understanding of what constituted measurement error; i.e., that error had many sources. This led to the 1951 paper "Coefficient Alpha and the Internal Structure of Tests." The subsequent results of this work led to further research and ultimately a rewriting of generalizability theory with Goldine Gleser. Referred to as G Theory, these results brought together mathematics and psychology as an aggregate structure within which error sources may be identified.

—Suzanne M. Grundy

See also Coefficient Alpha

Further Reading

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Cronbach, L. J. (1989). Lee J. Cronbach. In G. Lindzey (Ed.), *A history of psychology in autobiography, Vol. 8* (pp. 62–93). Stanford, CA: Stanford University Press.

Lee Cronbach's two most influential papers: <http://psychclassics.yorku.ca/Cronbach/construct.htm> and <http://psychclassics.yorku.ca/Cronbach/Disciplines>

CULTURE FAIR INTELLIGENCE TEST

Fluid intelligence (Gf) taps "the level of complexity of relationships which an individual can perceive and act upon when he doesn't have recourse to answers to such complex issues already stored in memory" (Cattell, 1971, p. 99). It is "concerned with basic processes of reasoning and other mental activities that

depend only minimally on learning and acculturation" (Carroll, 1993, p. 624). Therefore, tests of Gf have little informational content and require the ability to see complex relationships between simple elements like number and letter series, figure classification, figure analogies, spatial visualization, block designs, matrices, and so forth.

The Culture Fair Intelligence Test (CFIT) was designed by R. B. Cattell expressly as a nonverbal test to measure Gf. The CFIT comprises three scales: scale 1 for ages 4 to 8 as well as for mentally retarded adults, scale 2 for ages 8 to 13 as well as for average adults from the general population, and scale 3 for people of above-average intelligence. Scales 2 and 3 each have two parallel forms (A and B), which can be alternately used for retesting. The majority of these tests can be administered collectively, except some subtests from scale 1. The CFIT is highly speeded and requires detailed verbal instructions for administration.

Scale 1 includes eight subtests, but only half of them are really culturally fair. This scale cannot be recommended because some subtests must be administered individually, requiring complex instructions, and effects of familiarity with language and the habits of listening and attending are not minimized.

Scales 2 and 3 are quite similar, differing only in their difficulty level. They comprise four subtests: figure series (the individual is asked which figure logically continues a series of three model figures), figure classification (the individual is asked which two figures in each series go together), matrices (the individual must determine which of five alternatives most logically completes a given matrix pattern), and figure generalization (the individual must figure out the general rule for where a dot has to be placed by inferring the rule and picking the figure to which it applies). Each form of scales 2 and 3 takes about 30 minutes to administer.

Internal consistency and alternate-form reliability estimates generally range from .70 to more than .80. The lowest estimates are justified by Cattell on the scales' inclusion of several diverse formats to measure the same underlying construct. This is a main difference between the widely known Progressive Matrices Test and the CFIT. Cattell and Jensen

discredited the former because of specific variance due to using only the matrix-problem format. These researchers agree that the CFIT does not contaminate the measurement of the construct of interest (Gf) with variance specific to item type. A fine-grained measure of Gf must employ several different subtests to wash out any contamination with test specificity.

Validity has been researched by means of factor analysis and correlations with other tests. CFIT correlates around .80 with a latent *g* factor, and correlations with other reasoning tests (i.e., Raven's Progressive Matrices, the Wechsler Intelligence Scale for Children, and the Wechsler Adult Intelligence Scale) are generally above .50. Predictive validity studies show moderate correlations with several scholastic and occupational criteria.

The CFIT is an excellent choice for assessing intelligence across cultures because of its fluid nature. It has been administered in several European countries and in North America, Africa, and Asia, and norms tend to remain unchanged in relatively close cultures.

—Roberto Colom and Francisco J. Abad

See also Fagan Test of Infant Intelligence; Gf-Gc Theory of Intelligence; Intelligence Quotient; Intelligence Tests

Further Reading

- Buj, V. (1981). Average IQ values in various European countries. *Personality and Individual Differences*, 2, 169–170.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton-Mifflin.
- Cattell, R. B. (1980). They talk of some strict testing of us—pish. *Behavioral and Brain Sciences*, 3, 336–337.
- Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: North-Holland.
- Colom, R., & García-López, O. (2003). Secular gains in fluid intelligence: Evidence from the Culture Fair Intelligence Test. *Journal of Biosocial Science*, 35, 33–39.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.
- Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

CUMULATIVE FREQUENCY DISTRIBUTION

Once a frequency distribution has been created and the data are visually represented by means of a histogram or a frequency polygon, another option is to create a cumulative frequency distribution, or a visual representation of the cumulative frequency of occurrences by class intervals.

A cumulative frequency distribution is based on the same data as a frequency distribution, but with an added column (cumulative frequency), as shown in Table 1.

The cumulative frequency distribution begins by the creation of a new column labeled “Cumulative frequency.” Then, the frequency in each class interval is added to all the frequencies below it. For example, for the class interval of 19–24, there are 3 occurrences and none below it, so its cumulative frequency is 3. For the class interval of 25–29, there are 4 occurrences and 3 below it, for a total of 7 (4 + 3) occurrences in that class interval or below it. The last class interval (65–69) contains 1 occurrence, and there is a total of 49 occurrences at or below that class interval.

Once the cumulative frequency distribution is created, the data can be plotted just as they were for a histogram or a frequency polygon. Another name for a cumulative frequency polygon is an *ogive*. And if

Table 1 Cumulative Frequency Distribution

<i>Class Interval</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
65–69	1	40
60–64	4	39
55–59	5	35
50–54	4	30
45–49	3	26
40–44	5	23
35–39	6	18
30–34	5	12
25–29	4	7
19–24	3	3

the distribution of the data is normal, then the ogive represents what is popularly known as a bell curve or a normal distribution.

—Neil J. Salkind

See also Frequency Distribution; Histogram; Ogive

CURRICULUM-BASED MEASUREMENT

Curriculum-based measurement (CBM) is a method for measuring student competency and progress in the basic skill areas of reading fluency (e.g., words read correctly per minute), math computation, written expression, and spelling. When using CBM, an examiner gives the student brief, timed samples or *probes* lasting from 1 to 5 minutes, depending on the skill being measured, and student performance is scored for speed and accuracy to determine proficiency. Although CBM has been used in educational settings for more than 20 years, it is probably more familiar to special education teachers and school psychologists than to general education teachers and other professionals.

History of CBM

CBM began with the work of Stanley Deno and a number of doctoral students at the University of Minnesota in the late 1970s. Borrowing from the field of applied behavioral analysis, Deno and his team developed a measurement system that could efficiently produce monitoring data, could be displayed in graphic form, and would permit students' academic progress to be evaluated in only a few minutes.

CBM was examined in the 1970s with school-aged children with and without disabilities to assess its technical quality (e.g., reliability, validity) and practical utility (e.g., ease of administration). Following this development and validation phase, interest in CBM expanded because it provided an efficient alternative to expensive and time-consuming norm-referenced tests and was closely aligned with the curriculum. CBM has had the support of the U.S. Department of Education since the 1980s.

Examples of CBM

Examples of the application of CBM to reading, mathematics, spelling, and writing follow.

1. In reading, students read aloud for 1 minute from reading probes taken from basal reading series or from other reading probes designed with some control for grade-based readability. The number of words read correctly per minute is the metric of interest for evaluating oral reading fluency. In practice, three reading probes are given, and the middle score is reported. Another reading measure commonly used is the maze-reading task, a multiple-choice *cloze* technique in which students read grade-level reading passages in which every seventh word has been deleted and replaced by a blank; students are asked to fill in the blank by selecting one of three alternatives that appear beneath the blank. The measure is scored by counting the number of correct word choices per 5 minutes.

2. In mathematics, students write answers to computational problems. The math probes last from 2 to 5 minutes, depending on the type of skill assessed. The number of digits correct and incorrect for each probe is counted.

3. In spelling, the examiner dictates words at specified intervals of time for 2 minutes, and the number of correct letter sequences and words spelled correctly is counted.

4. In writing, the student is given a "story starter" (e.g., "Jill got a surprise package in the mail") and is asked to write a story within 3 minutes. The number of words written, spelled correctly, correct word sequences, or both are counted.

CBM is designed to identify students whose level and rate (slope) of performance are below those of the reference group. Thus, equal weight is given to skill level (low achievement) and to progress (slope), regardless of the type of skill assessed.

Key Features of CBM

CBM has three key features. It is dynamic, it does not preclude the use of other measures, and it is designed

to measure basic skills. As a dynamic measure, CBM is sensitive to the effects of instruction. When a student's learning improves as a result of a short-term intervention (i.e., 6 weeks), CBM is designed to detect this improvement whereas norm referenced tests may not be sensitive enough to do so. CBM has been described as an "academic thermometer." In the medical field, temperature is an efficient, important indicator of health. As an academic thermometer, CBM is an efficient, important indicator of academic health. For example, the number of words read correctly from text in 1 minute is an accurate yet efficient indicator of general reading achievement, including comprehension.

Although CBM assesses important skills, it does not measure all behaviors in an academic realm, and so it does not preclude the use of other specific measures. Finally, CBM is designed to assess basic skills necessary for success in later courses (e.g., social studies, science) that in turn are necessary for employment. Assessment of basic skills acquisition is particularly important for low-performing students through sixth grade and for the majority of special education students, who often continue to experience significant difficulty in basic skills throughout their schooling.

CBM and Its Use in a Problem-Solving Model

CBM is typically used in a problem-solving model that defines academic problems within a specific context and attempts to solve these problems. The problem-solving approach involves five steps: (a) problem identification, (b) problem certification, (c) exploration of solutions, (d) evaluation of solutions, and (e) problem solution. CBM is used in every step of this process to determine the severity of the problem, set goals, plan interventions, and evaluate the efficacy of these interventions. Within this framework, problems are defined as *situational*. A problem is conceived as a discrepancy between what is expected and what is occurring in a specific context. For instance, for a student experiencing reading difficulty, the problem would be defined in terms of how the student reads compared with a particular local standard (e.g., a community) rather than a national standard. In this approach, the need to assess situations or the effects of

interventions in addition to assessing the student is stressed.

An assessment approach like CBM often challenges researchers and clinicians to reconsider how children are assessed. Instead of measuring a student's skills at a single point in time, the child is assessed on an ongoing basis to evaluate the effect of instructional interventions. In this way, CBM takes into account the instructional context, something often overlooked in traditional assessment approaches. CBM shifts the focus from a summative evaluation perspective (e.g., determining what a student has learned) to a formative evaluation perspective (e.g., determining what the student is learning) that is continuous during instruction. CBM data is collected on a repeated, frequent basis, rather than in simple pre- or posttest fashion, to help determine the conditions under which student progress is facilitated. If student response to an academic intervention is unsatisfactory, the intervention is changed in some meaningful way with the aim of improving learning.

Figure 1 illustrates how CBM data are graphed to reflect learning level and slope. Each data point in Figure 1 represents the number of words a

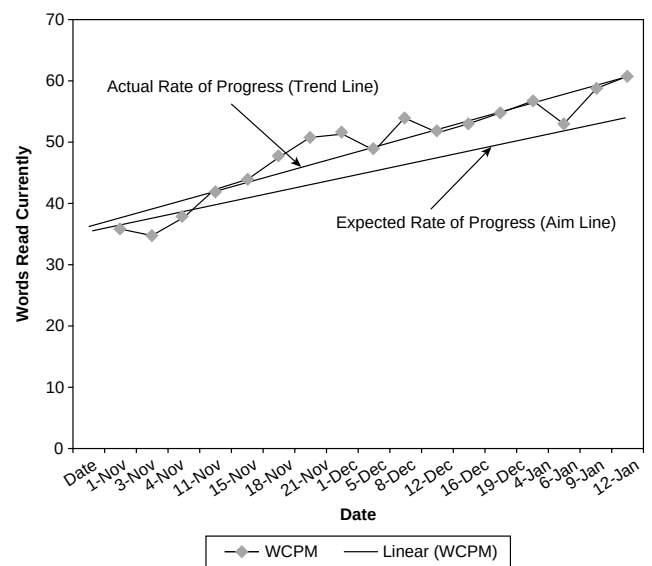


Figure 1 CBM Progress Monitoring Graph

Note: For example, the effects of an academic intervention on the reading achievement of a particular student over an 8-week period.

student read correctly per minute on a given day while exposed to a reading intervention. The heavy line indicates the trend, or the actual rate of progress under the academic intervention. The fine line is the aim line, or the expected level of performance for this child. The expected level of performance is an informed guess based on how typical students are expected to progress on a weekly basis. This information is used as a frame of reference for determining whether the referred student is performing at the same rate of improvement (i.e., slope) as classroom peers are. For example, there is some evidence to suggest that third-grade students typically improve about one word read correctly per week on a CBM oral reading measure. CBM practitioners call this learning trajectory the *average learning rate*. The average learning rate can help us make important comparisons among individuals (e.g., typical general education students and English language learners).

Advantages of CBM

Several advantages of CBM are evident. Stecker and Fuchs, for example, demonstrated improvements in academic growth among students with and without disabilities when teachers employed CBM, suggesting that CBM may be both an effective assessment tool and a powerful intervention technique to improve student performance. Because CBM takes only minutes to administer, it is possible to collect data on a large number of students and make local normative data available quickly to schools and districts. For certain populations, local norms may be preferred over national norms to identify which students are falling behind the general education curriculum. Local norms enable comparisons with students from similar geographic locations, who have similar curricular experiences and who share the same language or cultural background. CBM measures specific skills that are being taught in the classroom, ensuring content validity, whereas traditional norm-referenced tests often measure broad curriculum areas that may not be as well aligned with the curriculum. The alignment of test content and the curriculum plus the administration of multiple brief probes helps ensure valid, continuous

data about progress. Finally, recognition of CBM as a viable instrument to assess student performance can be found in a variety of task force documents on assessment and special education reform. CBM is also an integral part of the new, alternative approach to the identification of learning disabilities called Response to Intervention, which is part of the 2004 reauthorization of Individuals with Disabilities in Education Improvement Act.

Limitations of CBM

Despite the general acceptance of CBM as a measurement instrument, it has been subjected to criticism. Critics point out that CBM cannot assess the breadth and depth of the elementary, middle, and high school curriculum (e.g., content knowledge in various subjects). Advocates counter that CBM is intended to have an exclusive focus on basic skills assessment, not the assessment of broader skills or abilities. Critics also argue that a reliance on local norms would lead to variability between schools in the identification of students who are eligible for special education services under the Response to Intervention model, resulting in classification inconsistency from school to school. Proponents of CBM acknowledge this criticism but point out that classification variability already exists. Despite the criticisms, many followers of CBM would say that this measuring tool has earned a place among other measurement techniques that seek to assess academic difficulties and improve instruction within schools and districts.

—Romilia Domínguez de Ramírez
and Thomas Kubiszyn

Further Reading

- Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York: Guilford.
- Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford.
- Stecker, P., & Fuchs, L. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice*, 15, 128–135.

CBM testing materials page: <http://aimsweb.com>

Intervention Central: <http://www.interventioncentral.org>

CURSE OF DIMENSIONALITY

Curse of dimensionality refers to the rapid increase in volume associated with adding extra dimensions to a mathematical space. In the behavioral and social sciences, the mathematical space in question refers to the multidimensional space spanned by the set of V variables collected by the researcher. Simply put, the ability to simultaneously analyze large sets of variables requires large numbers of observations due to the fact that, as the number of variables increases, the multidimensional space becomes more and more sparse. This problem manifests itself in several analytical techniques (such as multiple regression and finite mixture modeling) in which difficulties arise because the variance-covariance matrix becomes singular (i.e., noninvertible) when the number of observations, N , exceeds the number of variables, V . Additionally, as N approaches V , the parameter estimates of the aforementioned models become increasingly unstable, causing statistical inference to become less precise.

For a mathematical example, consider multiple regression in which we are predicting y from a matrix of explanatory variables, $\mathbf{X}$. For ease of presentation, assume that the data are mean centered; then the unbiased estimate of the covariance matrix of $\mathbf{X}$ is given by

$$\Sigma = \mathbf{X}'\mathbf{X} \frac{1}{n-1}.$$

Furthermore, the general equation for multiple regression is

$$y = \mathbf{X}\beta + \varepsilon,$$

where

y is the $N \times 1$ vector of responses,

$\mathbf{X}$ is the $N \times V$ matrix of predictor variables,

β is the $V \times 1$ vector of parameter estimates corresponding to the predictor variables, and

ε is the $N \times 1$ vector of residuals.

It is well known that the estimate of β is given by

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y.$$

It is easily seen that the $(\mathbf{X}'\mathbf{X})^{-1}$ is proportional to the inverse of Σ . Thus, if there are any redundancies (i.e., Σ is not of full rank, or in regression terms, multicollinearity exists) in Σ , it will not be possible to take the inverse of Σ and, consequently, it will not be possible to estimate β . One possible introduction of multicollinearity into Σ is when V exceeds N .

Related to this general problem is the fact that, as V increases, the multidimensional space becomes more and more sparse. To illustrate, consider the Euclidean distance between any two points x and y ,

$$d(x, y) = \sqrt{\sum_{i=1}^V (x_i - y_i)^2},$$

the square root of the sum of squared differences across all V dimensions. To begin, consider the two points $x = (1, 3)$ and $y = (4, 7)$, which results in the Euclidean distance of $d(x, y) = [(1 - 4)^2 + (3 - 7)^2]^{1/2} = [9 + 16]^{1/2} = 5$. Now assume that K additional, albeit meaningless, dimensions are added to each observation by sampling from a uniform distribution with lower bound of 0 and upper bound of 1 (denoted by $U(0,1)$). The new Euclidean distance, $d(x, y)^*$, is given by

$$d(x, y)^* = 5 + \sqrt{\sum_{k=1}^K (U(0,1)_k - U(0,1)_k)^2},$$

where the 5 represents the original Euclidean distance and the remainder of $d(x, y)^*$ represents the additional distance that is due to random noise alone. Clearly, as $K \rightarrow \infty$, then $d(x, y)^* \rightarrow \infty$, indicating that as more dimensions are added, the two points become farther and farther apart. In the extreme, an infinite amount of random noise results in the two points being infinitely far apart.

The problem is amplified when considering the computation of multivariate distance, D (also known as the Mahalanobis distance), between two $V \times 1$ vectors, $\mathbf{a}$ and $\mathbf{b}$:

$$D(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})'\Sigma^{-1}(\mathbf{a} - \mathbf{b}).$$

If Σ equals the $V \times V$ identity matrix, then $D(\mathbf{a}, \mathbf{b})$ reduces to $d(\mathbf{a}, \mathbf{b})$, and the aforementioned problem exists when dimensions are added; however, when Σ does not equal the identity matrix, the aforementioned problems are accompanied by the fact that N must exceed V so that the inverse of Σ may be computed. Thus, any multivariate statistical technique (for example, multiple discriminant analysis) that relies on the inverse of the covariance matrix is subject to the curse of dimensionality.

To make the example more salient, consider the following visual demonstration. First, we generate 100 observations from a $U(0,1)$ distribution in one dimension (see upper left panel of Figure 1). Here the average distance between all points is .3415. The upper right panel of Figure 1 depicts the addition of an additional dimension, in which 100 more observations were generated from a $U(0,1)$ distribution, causing the average distance between all pairs of points to increase to .5436. The lower left panel of Figure 1 depicts the addition of yet another dimension generated from a $U(0,1)$ distribution, causing the average distance between all points to

increase to .6818. The lower right panel indicates the distance of all points to the mean in each of the scenarios. It is easily seen that as the number of dimensions increases, so do the distances to the mean.

This depiction illustrates how quickly point clouds can become sparse, wreaking havoc on data analysis that takes advantage of the full amount of information available to the researcher. In fact, infinitely increasing the number of dimensions also produces a multidimensional space with an infinite amount of variance. Instead of the traditional formula computing the covariance between two variables, x and y ,

$$\text{Cov}(x, y) = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) / (N - 1),$$

which is represented by the distance of observations from the mean, the covariance matrix can be represented as a set of pairwise distances,

$$\text{Cov}(x, y) = \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)(y_i - y_j) / (2N).$$

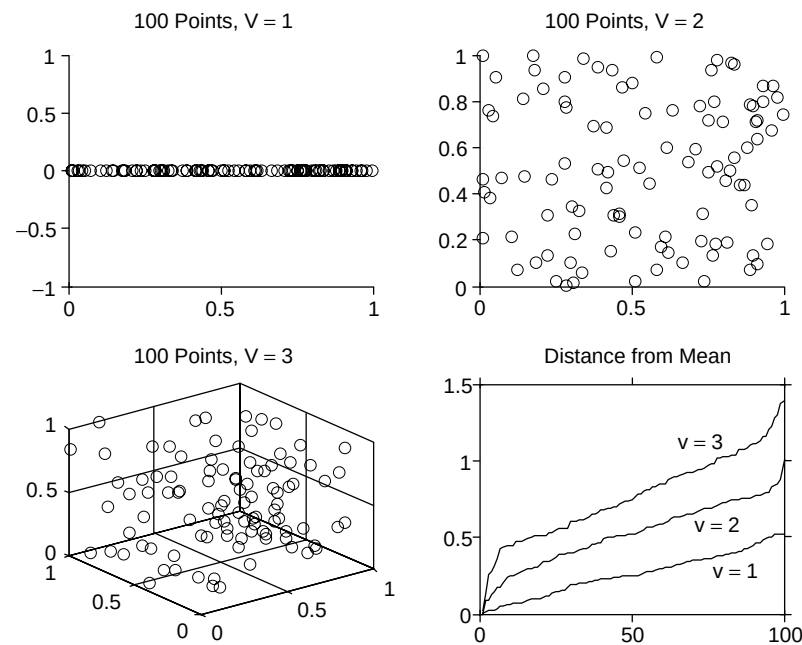


Figure 1 Interpoint Distance as a Function of Dimensionality

Given the result discussed above, that an infinite number of dimensions results in an infinite distance between any pair of points, it is clear that the variances or covariances will also approach infinity as the number of dimensions approaches infinity. This increase in variance leads to the unstable estimation of parameters and joint distributions in high-dimensional spaces.

In response to sparse high-dimensional spaces, social science researchers often attempt to reduce the dimensionality of the data in such a manner as to retain the relevant information in the full dimensionality in substantially fewer dimensions. The two most popular techniques of data reduction are principal components

analysis and factor analysis. Principal components analysis attempts to extract a set of K new variables—called components—(where K is less than V) from the original data such that the K components are linear combinations of the original V variables. Thus, each of the new variables contains some information from each of the original variables. Furthermore, the K components are mutually orthogonal (i.e., uncorrelated). Usually, K is chosen such that an adequate amount of the variance originally present in V is still explained in the reduced dimensionality. In general, principal components analysis is a pure data reduction technique. On the other hand, factor analysis is the more popular method employed in the social sciences. Similar to the components in principal components analysis, the factors derived from factor analysis are linear combinations of the original V variables. However, unlike principal components analysis, factor analysis assumes that an underlying model consisting of a small number of factors (in comparison with the original number of variables) gives rise to the observed set of V variables, where the goal is to recreate the original $V \times V$ covariance matrix with a substantially smaller number of factors. After conducting a principal components analysis/factor analysis, the researcher often uses the components or factors in subsequent analyses (such as regression, MANOVA, or cluster analysis). Although this is a reasonable approach to dealing with the curse of dimensionality, the researcher should always remember that some information is necessarily lost when the data are reduced to a lower-dimensional space and that appropriate cautions should be taken when making subsequent inferences and interpretations.

—Douglas Steinley

Further Reading

- Bellman, R. E. (1961). *Adaptive control processes*. Princeton, NJ: Princeton University Press.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.

CURVILINEAR REGRESSION

Researchers often use regression techniques to describe the relationship between two (or more) variables. In the simplest case (bivariate linear regression), it is assumed that the relationship can be described well by a straight line, $Y = a + bX$. One can use Student t to test hypotheses about or construct confidence intervals around the regression coefficients a (intercept or constant) and b (slope, number of units Y changes for each one-point change in X).

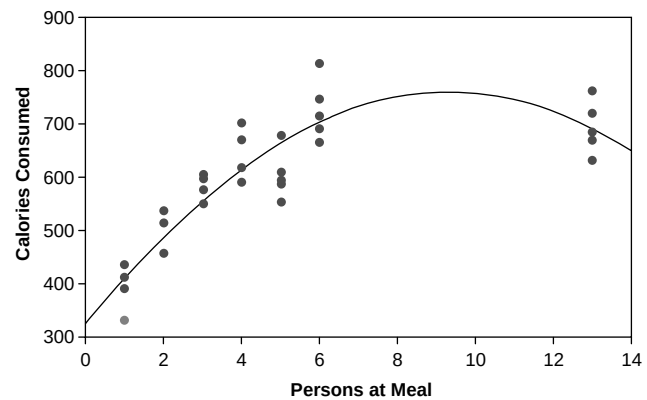
Often the relationship between variables can be better described with a line that is not straight. Curvilinear regression can be employed to describe some such relationships. In some cases, the researcher has good reason to expect a particular curvilinear relationship even before the data are collected. For example, an microbiologist may expect that the relationship between elapsed time and the number of bacteria in a rich medium is exponential. A psychophysicist may expect that the perceived intensity of a visual stimulus is a function of the logarithm of the physical intensity of the stimulus. A psychologist may expect that the relationship between the amount eaten by an individual and the number of persons present at the meal is a power function.

In other cases, the researcher does not expect any particular curvilinear relationship but discovers during data screening that the variables are not related in a linear fashion. One should always inspect a scatter plot of the data before conducting a regression analysis. All too often, researchers employ linear regression analysis when a much better fit would be obtained with a curvilinear analysis. Most researchers are familiar with linear regression and wish to stay within that framework when dealing with curvilinear relationships. This can be accomplished by applying a nonlinear transformation to one or both variables and then conducting linear regression analysis with the transformed data. Alternatively, one can conduct a polynomial regression (also known as a trend analysis), which is a multiple linear regression in which powers of the predictor variable(s) are included in the

Table 1 Simulated Data to Illustrate Curvilinear Regression

<i>Number of Persons at Meal</i>	<i>Calories Consumed by Individual</i>
1	413
1	332
1	391
1	392
1	436
2	457
2	534
2	457
2	514
2	537
3	551
3	605
3	601
3	598
3	577
4	701
4	671
4	617
4	590
4	592
5	587
5	596
5	611
5	552
5	679
6	745
6	692
6	666
6	716
6	815
13	762
13	631
13	670
13	720
13	685

model. Less commonly, the researcher may choose to conduct a true nonlinear analysis. In such an analysis, the researcher estimates the values of the model parameters by attempting to minimize the squared residuals (differences between observed Y and predicted Y) for the actual nonlinear model rather than attempting

**Figure 1** Scatter Plot With Quadratic Regression Line

to minimize the squared residuals for a linear regression on transformed data.

The Case Study and the Data

J. M. de Castro and E. M. Brewer investigated the relationship between the amount eaten by individuals and the number of persons at the meal. Table 1 presents data produced by a simulator that was designed to produce data similar to those obtained by de Castro and Brewer.

Plotting the Data

SPSS was employed to create a scatter plot with a quadratic regression line (Figure 1). Clearly, the relationship between number of persons and amount eaten can be described with a curved line better than with a straight line.

Polynomial Regression

For a linear regression, we would find a and b such that the errors in prediction would be minimized for the model $C = a + b_1P$, where C is calories consumed by the individual, and P is the number of persons at the meal. For a quadratic regression, we would add persons squared to the linear model, that is, $C = a + b_1P + b_2P^2$. Adding P^2 to the model allows the curve to have one bend in it. For a cubic regression, we would add

persons cubed to the quadratic model, that is, $C = a + b_1P + b_2P^2 + b_3P^3$. Adding P^3 to the model allows a second bend in the regression line. For each additional power of the predictor variable added to the model, an additional bend is allowed in the regression line.

While it is not impossible to fit the model to the data with hand computations, it is extremely laborious and shall not be done here. Any of the major statistical programs can do the analysis with the same routine that is used to conduct multiple linear regression. Here is an SAS program that will conduct the polynomial regression:

```
data eat; infile 'C:\CurviData-Sage.txt';

input persons calories; pers2 =
persons*persons; pers3 = persons**3;

proc reg; LINEAR: model calories =
persons;

QUADRATIC: model calories = persons pers2;

CUBIC: model calories = persons pers2 pers3; run;
```

The statistical output for the linear model shows that there is a significant linear relationship between calories consumed and the number of people at the meal, $C = 489.4 + 20.97 \cdot P$, $r^2 = .464$, $p < .001$. If we had not produced and inspected a scatter plot, we might be tempted to report this linear analysis. Look at the statistical output for the quadratic model (Figure 2).

The relationship is now estimated as $C = 322.6 + 93.63 \cdot P - 5.012 \cdot P^2$. Notice that the proportion of variance explained by the model has increased from .464 to .812. This increase in R^2 is statistically significant, $t(32) = 7.69$, $p < .001$. Allowing one bend in the regression line produced significantly better fit to the data.

The R^2 for the cubic model (.816) is only slightly greater than that for the quadratic model, and the increase in R^2 falls well short of statistical

**The REG Procedure Model:
QUADRATIC Dependent Variable: Calories**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	364763	182382	69.02	<.0001
Error	32	84559	2642.46099		
Corrected Total	34	449322			
Root MSE		51.40487	R-Square		0.8118
Dependent Mean		591.22857	Adj R-Sq		0.8000
Coeff Var		8.69459			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	322.61068	26.02751	12.39	<.0001
persons	1	93.63192	9.74149	9.61	<.0001
pers2	1	-5.01215	0.65194	-7.69	<.0001

Figure 2 SAS Output for Quadratic Model

significance, $t(31) = 0.86$, $p = .40$. Accordingly, the quadratic model is adopted.

Evaluating a Power Function

De Castro and Brewer tested the hypothesis that amount eaten (C , kilocalories) by individuals is a power function of the number of persons (P) at the meal. That is, $C = k \cdot \log(P)$. Such a model can be estimated by applying a log transformation to both variables and then using a simple linear model to predict one transformed variable from the other transformed variable. Here is the SAS code to conduct such an analysis for our simulated data:

```
data power; set eat; logpers = log10(persons); logcal =
log10(calories);

proc reg; POWER: model logcal = logpers; run;
```

The resulting model is statistically significant, $r^2 = .748$, $t(33) = 9.91$, $p < .001$. The prediction equation, in log terms, is $\log(C) = 2.629 + .237 \cdot \log(P)$. Exponentiating both sides of this expression transforms

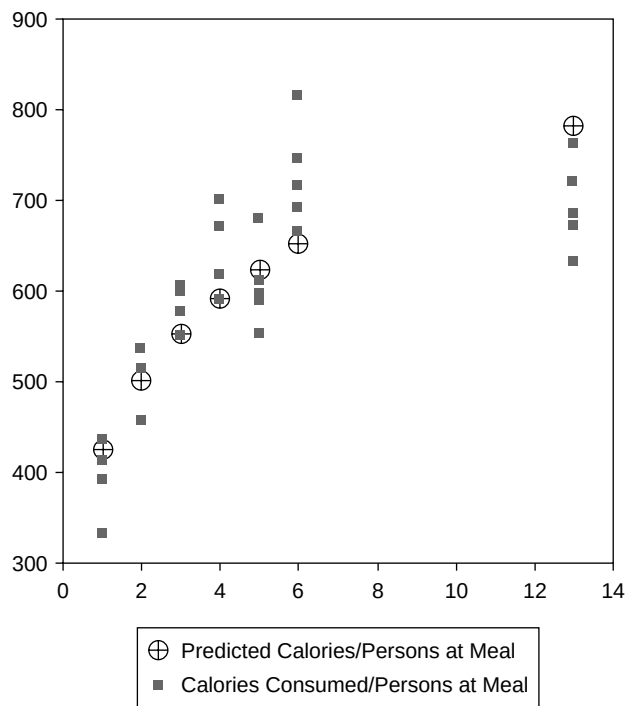


Figure 3 Scatter Plot With Values Predicted by the Power Function

the prediction equation to $C = 425.6 \cdot P^{.237}$. Figure 3 shows the data with the values predicted by the power function overlaid.

Confidence Interval on R^2

It is desirable to report effect size estimates and to place confidence intervals around those estimates. In regression analysis, the squared correlation coefficient is commonly used as an effect size estimate. Constructing a confidence interval around r^2 requires an iterative procedure, but there are scripts and programs that make this easy. If the predictor variable(s) can be considered to be fixed rather than random, use the SAS program at <http://core.ecu.edu/psyc/wuenschk/SAS/Conf-Interval-R2-Regr.sas> or the SPSS script at <http://core.ecu.edu/psyc/wuenschk/SPSS/CI-R2-SPSS.zip>. For our sample data, the predictor is most reasonably considered random. Jim Steiger's R2 program (which can be downloaded at www.interchg.ubc.ca/steiger/r2.zip, manual at <http://www.interchg.ubc.ca/steiger/r2.pdf>) will construct such a

confidence interval. For our data, a 95% confidence interval for the power model r^2 extends from .55 to .86.

Assumptions of the Analysis

There are no assumptions for estimating the slope, the intercept, and the r^2 , but one needs to keep in mind that the fit of the model to the data will be poor if one has chosen an inappropriate model, such as a linear model when a quadratic model would fit the data much better. When one is testing hypotheses about or creating confidence intervals around these estimates, there are three basic assumptions, each involving the error term (the residuals, the difference between actual values of Y and predicted values of Y):

1. **Independence:** The amount of error for each observation is assumed to be independent of the amount of error for any other observation.
2. **Homoscedasticity** (also known as homogeneity of variance): It is assumed that the error variance is constant across all values of the X variable(s).
3. **Normality:** It is assumed that the error term is normally distributed at each level of the X variable(s).

These assumptions are most commonly checked by inspecting residual plots—that is, plots with error on the ordinate and X (or predicted Y) on the abscissa. If the assumptions have been seriously violated, then one cannot trust the p values and confidence intervals.

Nonlinear Regression

In curvilinear regression, one obtains a curved line by applying nonlinear transformations to one or more of the variables and then employing a linear model. Truly nonlinear regression involves no transformations. The analyst provides a nonlinear model and starting values for each parameter in the specified model. The statistical software then uses an iterative process to fit the model to the data. Here is an SAS program used to fit a power function to our sample data:

```
proc nlin; parameters alpha=0 beta=1; model calories =  
alpha*persons**beta; run;
```

The nonlinear analysis converged on the model $C = 448.3 * P^{.204}$. The relationship was statistically significant, $\eta^2 = .694$, $F(2, 33) = 1503$, $p < .001$.

—Karl L. Wuensch

See also Linear Regression; Logistic Regression Analysis; Regression Analysis

Further Reading

de Castro, J. M., & Brewer, E. M. (1991). The amount eaten in meals by humans is a power function of the number of people present. *Physiology & Behavior*, 51, 121–125.

Luo, S., & Klohnen, E. C. (2005). Assortative mating and marital quality in newlyweds: A couple-centered approach. *Journal of Personality & Social Psychology*, 88, 304–326.

Rotton, J., & Cohn, E. G. (2000). Violence is a curvilinear function of temperature in Dallas: A replication. *Journal of Personality & Social Psychology*, 78, 1074–1081.

Curvilinear bivariate regression: <http://core.ecu.edu/psyc/wuenschk/MV/multReg/Curvi.doc>

Curvilinear regression: http://www.vias.org/tmdatanaleng/cc_regress_curvilinear.html

Nonlinear regression in SAS: http://www.ats.ucla.edu/stat/sas/library/SASNLin_os.htm

Polynomial regression with Stata: <http://web.archive.org/web/20000310035117/http://www.gseis.ucla.edu/courses/ed230bc1/notes3/curve.html>

D

Applied research or basic research: Research is what I'm doing when I don't know what I'm doing.

—Werner Von Braun

DARWIN, CHARLES (1809–1882)

Charles R. Darwin was born on February 12, 1809. He first studied medicine at Edinburgh, and then ministry at Cambridge. His primary interests were, however, in natural history. Ironically, Darwin found his early academic experiences in natural history “incredibly dull. The sole effect . . . was the determination never as long as I lived to read a book on geology, or in any way study the science.” Clearly, later academic and field experience reversed Darwin’s early views. After graduating, Darwin became the naturalist on the HMS *Beagle* in 1831. It was on the ship’s 5-year voyage that he gathered much of the evidence that later formed the basis for his ideas on evolution.

Darwin’s most important intellectual contribution was his theory of evolution by natural selection, published in *On the Origin of Species by Means of Natural Selection* (1859). Darwin later published *The Descent of Man, and Selection in Relation to Sex* (1871), which focused on human evolution, sexual selection, and cognitive/behavioral characteristics in humans and other species. In *The Expression of the Emotions in Man and Animals* (1872), Darwin

focused on patterns and mechanisms of emotional expression in humans and other animals. Darwin died on April 19, 1882, and was buried in Westminster Abbey.

Darwin’s ideas about evolution, natural and sexual selection, and work on emotional expression emphasized the adaptive and functional aspects of structure, behavior, and even cognition. This generated developments in measurement in areas ranging from comparative psychology to later work on human emotional expression (e.g., Ekman’s coding system of facial displays). Because Darwin’s theory emphasized individual variations, it helped to generate work in measurement of human individual differences. This influence was most clearly expressed through the work of Darwin’s cousin Francis Galton. As Galton himself notes in an 1869 letter to Darwin, “I always think of you . . . as converts from barbarism think of the teacher who first relieved them from . . . superstition. . . . Your book . . . was the first to give me freedom of thought.” Galton referred to his own work as the “natural history of human faculty.” Galton’s work on testing led to a number of developments in statistics, and measures of human abilities, physical characteristics, and sensory acuities. Through Galton, Darwin indirectly affected many important figures in

the history of measurement and testing, including Cattell, Pearson, and Spearman. Darwin's work thus had important direct and indirect influences on the fields of measurement and statistics.

—Matthew J. Hertenstein and Kevin E. Moore

See also Galton, Sir Francis

Further Reading

Simpson, G. G. (1995). *The book of Darwin*. New York: Washington State Press.

Darwin: <http://pages.britishlibrary.net/charles.darwin/>

Galton: <http://www.mugu.com/galton/index.html>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Shermer, M. B. (2002). This view of science: Stephen Jay Gould as historian of science and scientific historian, popular scientist and scientific popularizer. *Social Studies of Science*, 32(4), 489–524.

This paper pays needed attention to the depth, scope, and importance of Stephen Jay Gould's role as historian and philosopher of science, and his use of popular science exposition to reinforce old knowledge and generate new. It presents the results of an extensive quantitative content analysis of Gould's 22 books, 101 book reviews, 479 scientific papers, and 300 natural history essays in terms of their subject matter and places special emphasis on the interaction between the subjects and themes, how Gould has used the history of science to reinforce his evolutionary theory, and how his philosophy of science has influenced both his evolutionary theory and his historiography. **Darwin** said (and Gould cites) that, "All observation must be for or against some view if it is to be of any service." Gould followed Darwin's advice throughout his career, including his extensive writings on the history and philosophy of science.

DATA ANALYSIS TOOLPAK

The Analysis ToolPak is an Excel add-in that offers a special set of tools for completing a wide range of

statistical analysis. If the Data Analysis item doesn't appear on the Tools menu, it needs to be installed.

The Data Analysis ToolPak offers tools in the following general categories:

ANOVA
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F Test Two Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t Test
z Test

Using the Data Analysis Tool

As an example, examine the use of the ToolPak for completing a *t* test between two paired samples of observations, pre- and postintervention. In Figure 1, you see the scores for the 10 cases' data in column A (the Pre condition) and the scores for the same 10 cases in column B (the Post condition). Finally, you can also see the t-Test: Paired Two Sample for Means dialog box output of the Data Analysis ToolPak showing the following:

- Mean and Variance—the means and variances for each group
- Observations—the number of observations
- Pearson Correlation—the Pearson correlation coefficient
- The Hypothesized Mean Difference
- df—the degrees of freedom (*df*) associated with the *t* value
- *t* Stat—the *t* statistic

	A	B	C	D	E	F
1	Pre	Post		t-Test: Paired Two Sample for Means		
2	9	5				
3	7	6			Pre	Post
4	8	7		Mean	7.500	5.200
5	6	6		Variance	1.167	2.178
6	7	2		Observations	10.000	10.000
7	8	5		Pearson Correlation	-0.279	
8	9	4		Hypothesized Mean Difference	0.000	
9	8	5		df	9.000	
10	7	5		t Stat	3.535	
11	6	7		P(T<=t) one-tail	0.003	
12				t Critical one-tail	1.833	
13				P(T<=t) two-tail	0.006	
14				t Critical two-tail	2.262	
15						

Figure 1 Using the Data Analysis ToolPak for Analyzing Dependent Means

- $P(T \leq t)$ one-tail—the probability that a value of t would be different from chance for a one-tailed test
- t Critical one-tail – the critical value for rejection of the null hypothesis for a one-tailed test
- $P(T \leq t)$ two-tail—the probability that a value of t would be different from chance for a two-tailed test
- t Critical two-tail—The critical value for rejection of the null hypothesis for a two-tailed test

—Neil J. Salkind

See also Excel Spreadsheet Functions; Spreadsheet Functions

Further Reading

Salkind, N. J. (2007). *Statistics for people who (think they) hate statistics: The Excel edition*. Thousand Oaks, CA: Sage.

DATA COLLECTION

The word *data* is the Latin plural of the word *datum*, which itself is the past participle of the verb *dare* (DAH-reh), meaning “to give.” So, it literally means “things given.” *Data* is often used as a singular noun in English—what we call an “uncountable,” or a “mass term,” like “water,” “energy,” “information,” and so on, although the *Oxford Advanced Learner’s Dictionary* states that there is uncertainty with “data” as to whether it is singular or plural, and both are acceptable. But careful writers only ever use it as

plural. Although data are useful to generate information, knowledge, and wisdom, they in themselves are not treated as information or knowledge. What, then, are data? Although the lexicon meaning of data is facts or information, data need not be facts or information. Data are subjective and objective human experiences, feelings, attitudes, beliefs, values, perceptions, views, opinions, judgments, and so on. They are also objective facts in the universe, interactions between human beings and objective facts, and human subjective construction of objects and facts, irrespective of their object and factual reality. Thus, some data

are readily available as “things given,” whereas some data need to be diligently discovered and collected with ethical considerations, depending upon the research problem, need, and the researcher.

Many of us naturally and generally collect data, make sense of them, and use the same for better living. In the research world, purposeful and systematic data collection is an important and essential activity. It is one of the significant elements or phases within the research design that is followed by the research problem formulation (objectives, hypotheses, research questions, concepts, and variables); the selection of research; and sampling methods. It is preceded by data analysis and interpretation, and report reporting phases. Because data collection occupies a crucial phase in the research design, no research can be conducted without data. To a significant extent, the quality and impact of research depends upon high-quality, accurate, and uncontaminated data. In view of the significance and relevance of the data collection process for researchers, it may be delineated and discussed by addressing the following questions: Why do researchers collect data? What are the types of data? What are the data collection methods? When should data be collected? What are the ethical issues in collecting data, and how should researchers deal with them? What factors are likely to affect the quality of data? How can researchers minimize the factors that are likely to negatively affect the quality of data?

The Necessity of Data

Because social needs, problems, and causes keep constantly changing, new data need to be collected to understand and address these emerging changes. Toward this, some researchers collect data to explore and gain an in-depth understanding of the phenomenon, whereas others do so to answer their bold research questions, to test or formulate new hypotheses, or to validate or falsify existing theories by refining casual relationships or discovering new ones. At the extreme, new data are also useful to destroy existing paradigms and erect new ones. Data are also needed to formulate, implement, and evaluate appropriate policies, programs, and products of government and nongovernmental organizations and corporations. They also can be used effectively to inform or educate people and organizations about new trends that are relevant to them. From the postmodern perspective, data also play an important role in demonstrating multiple realities.

Types of Data

The universe is filled with a huge amount of data, so it needs to be categorized broadly for the systematic conduct of research and synthesis of research outcomes. All of the available data may be classified broadly into primary and secondary data, and each of these in turn may be categorized into quantitative and qualitative data. Primary data are collected directly from the field by observing, interviewing, or administering a questionnaire. Secondary data are collected from already available sources (see examples in Table 1). Data that cannot be measured by assigning a value or by ordering them in ascending or descending order are generally considered qualitative data, and data that can be subjected to some kind of quantification or measurement are generally considered quantitative data. Furthermore, data may also be categorized as tangible and intangible (e.g., smell, air, unexpressed feelings or emotions). However, it may be noted that these categorizations or dichotomies are researchers’ creations. In reality, these exist together, and thus in research, all types of data need to be collected and diligently integrated, if they enhance the understanding of reality.

Table 1 Types of Data

<i>Data</i>	<i>Qualitative</i>	<i>Quantitative</i>
Primary	Field observations, narratives	Age, income, educational level
Secondary	Letters, diaries	Census, annual reports

Data Collection Methods

Researchers often employ a specific method or several data collection methods to collect data, such as observation, case study, questionnaire, interview, focus groups, rapid rural appraisal, and secondary data. Some of these methods overlap with others, and some are more popular than the others. Many of these data collection methods have different variations within them; for example, the observation method has been further delineated into structured, unstructured, participant, and nonparticipant observation, and the case study method into intrinsic, instrumental, and collective case studies. There are different types of questionnaires and ways of administering them (one to one, in groups, or through mail, including e-mail). Interviews have been classified into structured, semistructured, and unstructured, which may be organized through face-to-face, by telephone, or by any other electronic mode. The focus group also has several types, including group interviews, group discussion, nominal group, and so on. These methods are very important because it is through these methods that data are collected.

Generally, research methodology books discuss details on these methods. Nonetheless, it is crucial to note a few points on them. First, researchers need to carefully select a method or a combination of data collection methods in such a way that they capture reality appropriately and accurately in order to answer the research questions and achieve the research objectives. Inappropriate or incorrect selection of data collection methods results in incorrect and misleading outcomes that distort the reality. Second, after having selected the most appropriate data collection method(s), researchers need to develop adequate knowledge and

skills through training, practice, or some other relevant means to use the data collection methods effectively. These may include sharpening observation skills, memorizing, taking notes, constructing a questionnaire or an interview schedule, asking questions, listening, moderating, dealing with diversions and interruptions, and respecting respondents' privacy and self-determination. Third, it is important to be aware of the strengths and limitations of various data collection methods and where and when they can be best used. Fourth, we should be aware of and effectively use in moderation data collection means with which we are all gifted. These are our five sense perceptions: eyes (seeing/observing), ears (hearing/listening), nose (smell), tongue (taste), and skin (touch). Just as some data collection methods are more often used than others (e.g., questionnaire or interview schedule), we might have gotten accustomed to using some sense perceptions more intensively than the others (e.g., too much speaking, not enough listening; observing but not noting; or too much listening/carried away with the field or respondent without observing and speaking). These sensory perceptions need to be employed effectively to collect data rather than relying only on the data collection instruments. Finally, while collecting data through the chosen method(s), researchers should ponder the following questions to keep the data collection process on track.

- What am I trying to discover?
- Why have I chosen the methods (research, sampling, data collecting) I have chosen?
- Do these methods help or hinder my efforts toward understanding reality?
- Are there any alternative methods to understand the phenomenon I am trying to understand?
- Do these categories of methods make any sense in understanding the reality?

Resources for and Timeliness in Data Collection

Data collection is a resource-intense activity in terms of time, money, and other resources, more so in the case of primary data collection. Researchers need to liberally estimate time, budget, personnel, and other resources, and make arrangements for the same in

advance to ensure a smooth data collection process. Most important, timeliness is very important in data collection. Researchers need to approach respondents, whether individuals, families, groups, communities, or organizations, at a time that is convenient to them and they are available and willing to provide data. Another important aspect of timeliness is that researchers need to be in the field when the events occur so as to collect data in the natural setting, if the research issue/design requires such an approach. For example, field data on mass protests, mob behavior, village fares, or indigenous methods of harvesting cannot be collected whenever researchers desire to collect. They have to be timely in collecting these types of data, just like a natural scientist can collect data on an eclipse only when it occurs.

Ethical Considerations

Researchers need to collect data according to the set ethical standards, which are often based on certain values and principles: honesty, truthfulness, privacy and confidentiality, self-determination and voluntary involvement, zero physical and psychological harm, dignity and worth of human beings, accountability, right to know on the part of respondents, fairness and impartiality on the part of researchers, and informed consent. On the other hand, researchers should avoid breach of confidence and agreements, absence of informed consent or self-determination/autonomy of respondents, deception, risk of harm or offense, acts involving conflict of interest, and any unethical act.

Many government and nongovernment organizations, universities, and research firms have well-developed research ethics committees and ethics clearance application forms. Before beginning the data collection process, researchers should adhere to these ethical requirements and collect data accordingly. Those researchers who do not belong to any organizations or whose organizations have not developed such ethical standards and requirements should also collect data by setting their own ethical standards based on the above stated values and principles. They should explain the nature and purpose of research, provide satisfactory answers to all questions, inure

that respondents are involved voluntarily and that no force is used, and allow the respondent to withdraw from the research at any time if he or she wishes to do so.

Impediments in Data Collection

Data collection is a planned, purposeful, and systematic activity. Despite choosing appropriate data collection methods; meticulously developing data collection instruments; planning adequate resources, including time; and meeting ethical standards, researchers may encounter several impediments in the data collection process. One probable reason for these impediments is that the nature of the setting, the research problem, the researcher, the researched, the time of research, and the prevailing social conditions vary every time. Thus, the data collection impediments may be analyzed by looking at three “R” factors: the researcher; the research problem; and the researched, or a combination of these factors.

Because the researcher is the main actor in the data collection process, he or she can contribute significantly to reducing or increasing field difficulties. Data collection experiences suggest that there are three main issues related to the researcher. First, researchers’ state of mind affects the data collection process because they may sometimes feel nervous, anxious, incapacitated, irritated, uncomfortable, overwhelmed, frightened, frustrated, tired, and at times less confident. Several factors within and outside the researcher may contribute to such a state that might affect researchers’ observation, interviewing, responding, and note-taking abilities. Second, researchers’ negative attitudes, prejudices, and preconceived notions toward the research problem, the field, respondents, and communities may interfere with the data collection process and reduce the quality of data. Finally, researchers’ action (i.e., how they actually behave in the field and with respondents) is also important and may obstruct the data collection process if not appropriate.

The second factor is the research problem. Some data collection difficulties are related to the nature of the research problem and the decision researchers

make to enter particular settings. If research problems deal with sensitive issues such as drug addiction, bankruptcy, the accused awaiting trial in the criminal justice system, ethnicity, development of toddlers and children, and so on, researchers often experience several challenges while collecting data. Data collection experiences have demonstrated that some respondents or communities may feel threatened and insecure because of the sensitivity of the issue. In some cases, data are simply not available, accessible, or discloseable. For example, while tracing genealogies of families, information on women may not be available in some cultures. In some regions and towns, it may not be possible to locate the universe of the community. Census reports may not have a particular type of information. A complete, up-to-date, and accessible list of agencies, organizations, and companies may not be available. At times, researchers may not have access to needed data or organizations. Information may not be well recorded and kept. These are real problems in the field that are beyond the control of researchers, and they can affect the quality of the data collection process. The difficult nature of the setting and lack of information about the setting (e.g., widely dispersed respondents or communities in rural, remote, and hilly areas; unclear addresses and road maps, etc.) may also lead to exhaustion and thereby weaken the data collection process, including its pace.

The third source of data collection impediments is the researched (i.e., respondents and communities). Data collection experiences suggest that researchers have faced the most common problem of making an entry (into the community) and gaining acceptance. Every means or way of approaching the respondent and the community (e.g., through written letters; health officials; government officials; local leaders, political or otherwise; friends/relatives; or independently without anybody’s introduction) has pros and cons and may affect the accuracy of data being collected. Equally important is gaining acceptance. If the respondent’s suspicions and doubts are not cleared, and acceptance is not gained, the data collection process will be hampered significantly, and that, in turn, may lead to inconsistent and incomplete data.

Experiences of interviewing respondents have revealed that an unsuitable location for the interview, lack of functional trust, refusal to give an interview, difficulty in convincing the respondents, interference by friends or members of the family, respondents' keenness to complete the interview quickly, more talkative respondents, and not knowing the local language can pose several impediments to the data collection process. In terms of the questionnaire, faulty design of the questionnaire, low return rates, difficulties in collecting a group of respondents at one place, lack of organizations' support to employees in completing the questionnaire, and approaching busy professionals at their workplace have hampered the data collection process. Ethical issues in observation studies, planned or arranged observations, and lack of prompt recording of observations appear to affect the quality of collected data. Delays in obtaining permissions to collect data from organizations, particularly from the government, and lack of cooperation of staff members to give access to the available data also create problems in data collection. Other factors such as adverse weather conditions, high sample mortality rates, lack of adequate resources, isolation, and health issues of the researcher also may get in the way of data collection.

Strategies to Ensure High-Quality Data Collection

Although the above presented impediments can affect the data collection process and reduce the quality of data, researchers can consciously employ some systematic strategies to ensure the collection of accurate data. In regard to the impediments stemming from the researcher, first, researchers need to be aware of their state of mind and reflect on it by raising the following questions: Why do I feel this way? What am I doing here? What are my attitudes toward respondents and communities? How am I behaving with people in the field? To what extent does my state of mind affect my data collection process? Is it blocking my efforts to understand field realities? How can I overcome these contextual feelings (state of mind) and change my attitude and behavior, if necessary? Second, these reflections should result in enhancing the competence

of researchers by acquiring needed knowledge, by developing practice skills and appropriate attitudes, and by taking right actions. It is important for researchers to feel comfortable and confident in the field, and enhanced competence will help achieve it. Finally, researchers' experiences suggest that additional reading, better information about the issue, adequate practice, acquaintance with the field, use of professional skills, anticipation of problems, and preparation of possible remedies will help. Regardless of respondents' background, status, communities and conditions, and cooperation or noncooperation, researchers should respect them. They also should be free from their own prejudices and preconceived notions about the field so as to develop conducive attitudes and behave appropriately in the field. In addition, researchers need to be assertive and flexible.

Several creative strategies need to be explored to prevent and to deal with data collection difficulties emanating from the research problem and setting. When the research issue is sensitive and respondents feel insecure and threatened, it is less likely that a good data collection process will begin. Strategies toward this issue will be discussed shortly. If the research problem and setting-related data collection difficulties are beyond the control of researchers, first, they should not get perturbed; second, they should study the problem; and third, they should look at possible alternatives. Once they analyze the possible alternatives, the most appropriate alternative can be chosen and changes can be introduced in the data collection strategies. Thorough pilot study should certainly signal such potential problems. Researchers need to anticipate and plan well, including logistics to cope with some of the realistic difficulties in the field. Careful use of local guides/volunteers and resources may reduce some of the problems. When research is undertaken in rural and remote communities and tribal areas, researchers must learn to live happily with limited facilities and without the luxuries of urban life. The pace of research work needs to be organized in such a way that it takes care of physical exhaustion. If it is not possible to collect data on some issues and from some settings, it may be necessary to alter the whole research design.

With regard to respondent-based data collection difficulties, a few strategies may be recommended. Because making an appropriate entry is a critical issue and there is no foolproof strategy to address it, researchers need to be conscious of how they are going to make an entry and how they will access respondents, and they need to make an assessment about likely implications on the quality of data. An analysis of the consequences of each entry option on data to be collected may be undertaken, and an entry approach that has minimum consequences on the data may be followed. It is also important to develop systematic plans to overcome those consequences. Another approach is that when the researcher feels confident that initial data were inconsistent and unreliable, such data may be excluded once the reliable data pattern is established. To gain acceptance and to deal with sensitive issues, researchers need to build functional trust and rapport, and establish credibility. Toward this, researchers need to provide simple, straight, and honest information to respondents, communities, and organizations, and answer all questions so as to overcome their suspicions and doubts. Efforts to overcome this problem might include ensuring direct contact with the respondent, rather than using a second person or intermediary to approach the respondent; maintaining strict confidentiality; suppressing actual names; exploring the respondent's version of the events, opinions, and so on; and avoiding using anything (e.g., tape recorder) that the respondent particularly finds threatening. Researchers should avoid defensive arguments with the respondents. They also must follow ethical guidelines that are appropriate to respondents' cultural practices. Most important, researchers should demonstrate warmth, empathy, friendliness, and pleasantness; show interest in what respondents say; and allow additional questions and discussion that may not be related to instruments and the research problem. These strategies are likely to facilitate a better data collection process to obtain rich, reliable, and valid data.

A mutually convenient location should be chosen for the data collection, whether it is an interview, administration of a questionnaire, or a focus group discussion. In the case of the respondent's refusal to

provide data, researchers should politely thank him or her and withdraw from the process. It is also important to anticipate a range of interruptions from people other than respondents (e.g., relatives, friends, etc.) and prepare well to minimize them. Researchers need to prepare and plan well to work with the language difficulty, if they do not know the local language. They need to learn and develop local basic vocabulary. Most important, they need to identify, train, and employ neutral interpreters (who do not take the side of the researcher or the researched) who do not affect respondents and their responses. Long and exhausting data collection instruments should be avoided. By pretesting, the optimum length should be estimated. If an instrument takes a long time, breaks should be planned at appropriate stages of the data collection. In-depth or long interviews may be conducted in two to three separate sessions. If particular items of the interview/questionnaire do not work, the researcher should be flexible enough to consistently drop them from the schedule.

Recording of data, whether through handwritten notes or electronic devices, should be avoided if it is implicitly or explicitly resisted by respondents. An overreliance on electronic gadgets is not recommended because they may not work when researchers need them the most. If the data collection is based on the researcher's memory, the researcher must expand his or her notes and then write down his or her memories immediately after interviews. Delay would cause memories to fade and thus the collected data as well.

If questionnaire respondents are located in government, nongovernment, or business organizations, researchers may ask the organization head to issue a cover letter advising the respective employees to cooperate with the survey. This approach may facilitate the data collection process in organizations. Avoid contacting professionals during their busy hours, and approach them according to their availability and convenience.

In the case of a questionnaire, administering, completing, and collecting it in one session will yield better return rates than giving a questionnaire to respondents and asking them to return it later. Researchers must have some autonomy in observing

so that they can get an adequate picture of the phenomenon being observed. Research experiences show that sometimes meaningful data may be collected through casual experiences, observations, and conversations. Researchers may not be able to capture such meaningful data when they approach respondents with a questionnaire/interview schedule in a formal way. If permission is required, it should be obtained well in advance. If the research topic is sensitive and securing permission is doubtful, the researcher may start work on the topic only after obtaining the permission. If high sample mortality is expected, researchers should plan for a larger sample size. They should also consciously plan opportunities to overcome the problem of isolation in the field. Modern communication technologies (e-mail, Internet chat, etc.) may also be used to achieve this purpose, if they are accessible. Finally, researchers need to take necessary steps to take care of themselves and to maintain good health.

Conclusion

As stated in the introduction, it may be reiterated that data collection activity is a crucial aspect of the research design. This entry has discussed the necessity of data collection, types of data, several data collection methods, resources required for and timeliness in data collection, ethical considerations, impediments, and strategies to ensure the collection of high-quality and accurate data. It may be noted that this discussion is neither comprehensive nor conclusive. The suggested strategies may work for some and not for others. However, this entry may provide important leads to researchers to further explore data collection methods, impediments, and strategies.

—Manohar Pawar

See also Descriptive Research; Variable

Further Reading

- Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge, UK: Cambridge University Press.
- Cowie, A. P. (1989). *Oxford advanced learner's dictionary*. Oxford, UK: Oxford University Press.

- Kuhn, T. (1962). *The structure of scientific revolution*. Chicago: University of Chicago Press.
- Kuhn, T. (1974). *The structure of scientific revolution* (2nd ed.). Chicago: University of Chicago Press.
- Lee, R. M., & Renzetti, C. M. (1993). The problem of researching sensitive topics: An overview and introduction. In C. M. Renzetti & R. M. Lee (Eds.), *Researching sensitive topics*. Newbury Park, CA: Sage.
- Pawar, M. (2004). *Data collecting methods and experiences: A guide for social researchers*. Chicago: New Dawn Press.
- Pawar, M. (2004). Learning from data collecting methods and experiences: Moving closer to reality. In M. Pawar, *Data collecting methods and experiences: A guide to social researchers*. Chicago: New Dawn Press.
- Popper, K. (1965). *The logic of scientific discovery*. New York: Harper & Row.
- Smith, C. D., & Carolyn, D. (1996). *In the field: Readings on the field research experience* (2nd ed.). Westport, CT: Praeger.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Shields, C. M. (2003). Giving voice to students: Using the Internet for data collection. *Qualitative Research*, 3(3), 397–414.

Good data collection techniques are essential for a research project to run smoothly and for the data to be trusted. This article explores the use of a Web-based survey as a means of **data collection** with more than 450 adolescents in an American school district with approximately 50 percent visible ethnic minority students. After describing the context of the study, the author explores issues related to the ease of data collection, the potential challenges and promise of the Web-based format, and the quantity and quality of data collected. Carolyn Shields demonstrates that the data collected were extremely rich, and that students appeared to be more comfortable with the electronic data collection than with an in-person interview. Moreover, the inherent issues of power differential related to race, class, and position may be overcome using this strategy for data collection.

DATA COMPRESSION

Data compression is the process by which statistical structure in data is used to obtain a compact

representation for the data. Structure can exist in data in various ways. If there is a correlation between neighboring symbols, this correlation can be used to remove the predictable portion of the data and encode only what remains. If patterns exist in the data, they can be replaced by indices to a dictionary of patterns. Even when samples of a data sequence are independent of each other, they might show bias, with some symbols occurring more often than other symbols. This bias can also be used to provide compression. Sometimes, it is easier to focus on what is not present rather than what is present in the data. For example, the low pass nature of particular data can be taken advantage of by processing the data in the spectral domain and discarding the higher frequency coefficients. In brief, the characteristics of the data guide the compression process.

Depending on the requirements of the user, data compression techniques can be classified as lossless or lossy. Lossless data compression techniques allow the exact recovery of the original. Lossy data compression permits the introduction of distortion in a controlled fashion to provide greater compression. Lossy techniques are used only in situations where the user can tolerate distortion. We will discuss some commonly used data compression techniques in the following sections.

Application Areas

Data compression is used in a wide variety of applications. WinZip and Gzip are commonly used file compression utilities on computers. Images on the Internet and in many cameras are compressed using the JPEG algorithm. Video conferencing is conducted using compressed video. Cell phones use compression techniques to provide service under limitation of bandwidth. Digital television broadcasts would not be feasible without compression. In fact, compression is the enabling technology for the multimedia revolution.

Compression Approaches

Compression can be viewed, and compression techniques classified, in terms of the models used in the compression process and how those models are

obtained. We can focus on the data, examining the different kinds of structures that exist in the data without reference to the source of the data. We will call these approaches data modeling approaches. We can try to understand how the data are generated and exploit the source model for the development of data compression algorithms. Finally, we can examine the properties of the data user because these properties will impose certain constraints on the data. We begin by looking at techniques based on properties gleaned from the data.

Data Modeling Approaches

With different applications, we get different kinds of structure in the data that can be used by the compression algorithm. The simplest form of structure occurs when there is no symbol-to-symbol dependence; however, the data symbols take on different values with differing probabilities. Compression schemes that make use of this statistical skew include Huffman coding and arithmetic coding.

Huffman coding, developed as a class project by David Huffman, assigns short codewords to symbols occurring more often and long codewords to symbols that occur less often. Let's look at the example in Table 1. There are five symbols in the original file. If we were to represent them using a fixed-length code, we would need three binary digits to represent each symbol. However, if we assign codewords of different lengths to each symbol according to their probability, as shown in Table 1, the average length (l) of binary bits needed to represent a symbol will be

$$l = 0.5 \times 1 + 0.2 \times 2 + 0.15 \times 3 + 0.1 \times 4 + 0.05 \times 4 = 1.95 \text{ bits/symbol.}$$

Table 1 Huffman Coding

<i>Symbols</i>	<i>Probability</i>	<i>Binary Representation</i>	<i>Codeword</i>
A	0.5	000	1
B	0.2	001	01
C	0.15	010	000
D	0.1	011	0010
E	0.05	100	0011

Thus, on average, we save 1.05 bits ($3 - 1.95$) per symbol using Huffman coding to represent the original symbols. This might not seem like much until we consider the possibility that the source may be generating many millions of symbols per second, in which case the savings is on the order of millions of bits per second. The encoded sequence is uniquely decodable. By this we mean that a sequence of codewords corresponds to one and only one sequence of letters. This is because in a Huffman code, no codeword is a prefix of another. For example, if we see the bits 01, we have to decode it as B because no other codeword begins with 01. Thus, the “10100000100011,” will be parsed into “1 01 000 0010 0011” and will be decoded into “ABCDE.”

Another method for coding sequences in which some symbols occur with higher probability than others is arithmetic coding. In arithmetic coding, every sequence is assigned a subinterval in the unit interval $[0,1)$ where the size of the subinterval is proportional to the probability of occurrence of the symbol. The binary representation of a “tag” in this subinterval is truncated to the ceiling function of $\lceil \log_2(1/p) \rceil$ bits, where p is the probability of the sequence. Clearly, sequences that are more probable will have a shorter codeword than sequences that are less probable.

The reasoning behind using shorter codewords for more probable symbols can be extended to collection of symbols or phrases. Commonly occurring phrases in a text can be collected in a dictionary and encoded with an index into the dictionary. The problem then becomes one of constructing the dictionary. In two landmark papers in 1977 and 1978, Jacob Ziv and Abraham Lempel provided two different approaches to forming a dictionary. In their 1977 approach, the dictionary was simply a portion of the previously encoded text. A repeat of a phrase or pattern was encoded by sending the offset from the current text and the length of the text to be copied. This compression approach, known as LZ77, has been incorporated in such popular packages as gzip and PNG.

The second approach, proposed in their 1978 paper, built an explicit dictionary based on the past. Terry Welch popularized a variation of this approach, known as LZW, in a 1982 paper. Suppose we have the following input sequence: “HO_HO_HO_OPS_OPS_OPS.” Assuming that the

alphabet for the input file is {H, O, _, P, S}, the LZW dictionary initially looks like Table 2. The LZW algorithm finds the longest match in the dictionary to the sequence being encoded, encodes the index of this match, and concatenates the match with the next symbol to form a new entry in the dictionary. For the example sequence, initially, the beginning pattern “H” is the longest pattern in the dictionary, which is encoded as 1. At this time, “HO” is added as the sixth entry in the dictionary. Now, the uncoded sequence starts from “O” and the longest pattern in the dictionary will be “O,” which is encoded as 2. “O_” will be the seventh entry in the dictionary.

Table 2 Initial LZW Dictionary

<i>Index</i>	<i>Entry</i>
1	H
2	O
3	_
4	P
5	S

Continuing in this manner, we build the two-letter patterns in the dictionary. When we reach the letter “H” in the second “HO_,” we have eight entries in Table 3. The output sequence is 123. Now we can find matches of two letters in the dictionary and begin to construct patterns of three letters. The next match is “HO,” which is encoded as 6. The ninth entry “HO_” is added to the dictionary. Continuing in this manner, we will have the dictionary shown in Table 3 and the output sequence is 1 2 3 6 8 7 2 4 5 3 12 14 16.

Table 3 The LZW Dictionary for the Above Sequence

<i>Index</i>	<i>Entry</i>	<i>Index</i>	<i>Entry</i>
1	H	10	_HO
2	O	11	O_O
3	_	12	OP
4	P	13	PS
5	S	14	S_
6	HO	15	_OP
7	O_	16	OPS
8	_H	17	S_O
9	HO_		

Note that when the encoder first meets the “OPS” pattern, it begins to build the two-letter entries in the dictionary again in Table 3 (Index 12). The dictionary is built dynamically according to the content in the data file. The decoding is similar to the encoding process. Because the decoder does not know the symbol right after the decoded symbol, it begins to build the dictionary entry only after decoding the second symbol. So the dictionary build-up in the decoder is one step behind the encoder. As long as the encoder does not encode the pattern using the entry just put in, there will not be a problem. The LZW algorithm has the special handling procedure for using the most recent entry in the encoder, which we will not discuss here.

Modeling the Source of Information

The use of models of the source of information represented by the data is a successful approach to compression. In particular, this approach is used for the compression of speech signals before they are transmitted over the digital cellular network. The approach used in cell phones essentially involves generating a model for the speech to be compressed and transmitting the parameters of the model to the receiver along with some information about aspects of the speech not incorporated in the model. The receiver then regenerates the speech. Speech is produced by forcing air first through the vocal cords, then through the laryngeal, oral, nasal, and pharyngeal passages, and finally through the mouth and the nasal cavity. Everything past the vocal cords is referred to as the vocal tract. The vocal tract can be modeled by a filter. The vibration of the vocal cords can be simulated by pulse sequences, which are called excitation signals. The Code Excited Linear Predictive (CELP) algorithm, which is widely used in the cellular system, is based on this model of human

speech production to produce high-quality speech at low bit rates.

A diagram of a CELP encoder and decoder is shown in Figure 1. The input speech is divided into frames. For each frame, the parameters of the vocal tract filter are obtained from the original speech. An excitation signal is chosen from a stored excitation codebook. Because the human ear is very sensitive to pitch errors, a pitch filter is added between the excitation and the vocal tract filter, as shown in Figure 1. The index of the excitation signal, the gain of the excitation signal, and the parameters of the pitch filter are selected to reduce the perceptual difference between the original speech and the synthesized speech. After finding the best parameters, the encoder sends these parameters to the decoder instead of the original speech signal. At the decoder, the speech is synthesized based on these parameters.

Modeling the Users of Information

Examining the users of a class of information can tell us a lot about the characteristics of the information. The limitations of the users can provide opportunities for discarding information not perceptible to the user, thus leading to compression. Schemes that use this

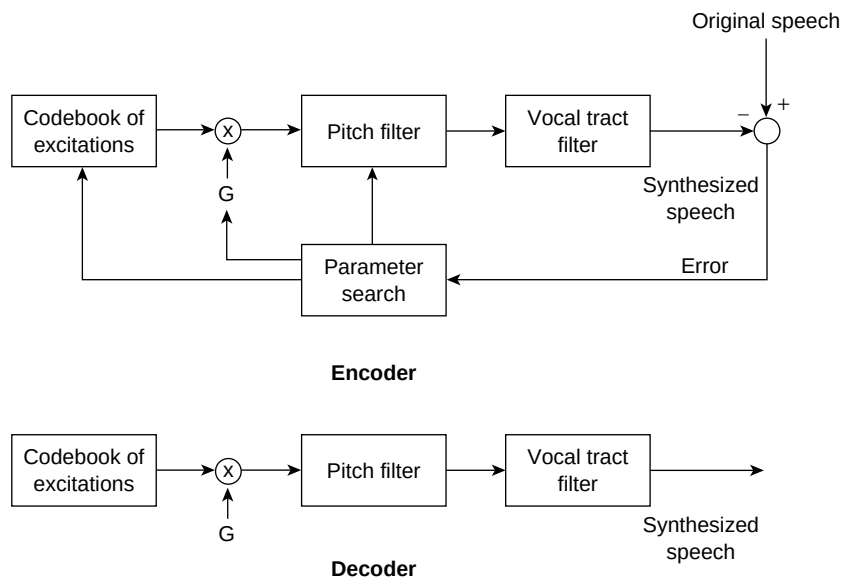


Figure 1 Diagram of CELP Encoder and Decoder

approach are particularly useful for classes of data that are not generated by a single type of source. For example, unlike human speech, music signals are generated by a variety of sources. Wonderful sounds can come from musical instruments, such as violins, pianos, and flutes, and can also be recorded from natural sources, such as birds, wind, or running water. It would be almost impossible to find a source model for these very different sources. However, although the sources are different, the user (of interest to us) is the human auditory system. Therefore, compression techniques can take advantage of the characteristic of the human auditory system.

The human ear can hear sounds from approximately 20 Hz to 20 kHz. However, even in this range, sounds below an audibility threshold cannot be heard. Furthermore, this audibility threshold can be raised locally through a phenomenon called *masking*. Spectral masking results in the raising of the audibility threshold in the spectral vicinity of a tone. Temporal masking results in the temporary raising of the threshold for a short period prior to and after a sound. Many audio compression algorithms are based on the masking effect. A block diagram of the popular MP3 algorithm is shown in Figure 2.

There are two important parts in the MP3 encoder—the filter bank and the psychoacoustic model. The psychoacoustic model finds the major components in the audio signal and calculates the masking curve imposed by these major components. The filter bank converts the audio input samples into samples of different frequency bands. The encoder then determines the active frequency bands and the quantization level needed for that band according to the given output bit rate and masking curve. For example, if the sound pressure level of the samples in a certain band is below the masking curve, the samples in this band are not needed because they are

not audible to humans. If the sound pressure level in a band is much higher than the masking curve, more quantization levels are needed for this band. In the final step, the encoder packs all this information into an MP3 bitstream.

The structure of the decoder is rather simple, as shown in Figure 3. The input stream is first unpacked; then the samples in each band are reconstructed; and the samples in all the bands are remapped into audio samples, which sound the same as the input audio samples.

As in the case of music, images and video are also generated by a variety of sources. Again, as in the case of music, although the sources are diverse, there is only a single user of interest to us, namely, the human visual system. The images of interest to humans consist of regions of constant or stationary pixel values. In other words, most of the image consists of regions of low spatial frequency. Therefore, we can design compression schemes that take advantage of this fact. An example of a scheme that does that is the popular JPEG algorithm. A block diagram of the JPEG compression algorithm is shown in Figure 4.

The input image is divided into 8×8 blocks. The Discrete Cosine Transform (DCT) transform is then applied to each block. This results in a spatial frequency representation of the block. The basis functions of the

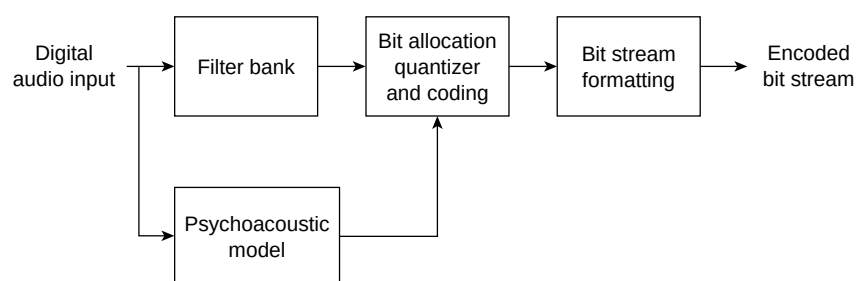


Figure 2 Block Diagram for the MP3 Encoder

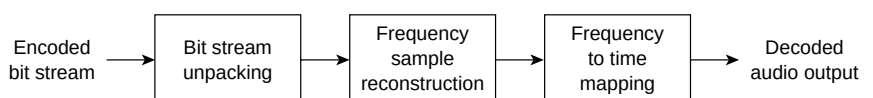


Figure 3 An MP3 Decoder

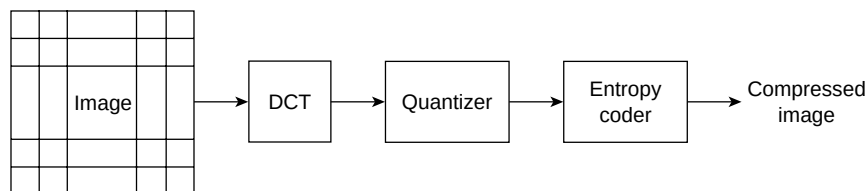


Figure 4 System Diagram of Baseline JPEG

DCT transform are shown in Figure 5. The low-frequency components are more important than the high-frequency components to characterize the images. This allows us to drop high-frequency components without degrading the perceptual quality of the reconstructed image. For example, Figure 6 shows the original image and reconstructed images obtained when some of the coefficients are discarded. The reconstructed images are based on only the first, second, and sixty-fourth DCT coefficients. It can be seen that the first few DCT coefficients contain most of the information about the image, whereas the sixty-fourth DCT coefficient has little visible information.

After the transform, the JPEG encoder quantizes these DCT coefficients. Because the coefficients with lower indices are more important, they are represented at a high resolution, whereas the coefficients with

higher indices are represented more crudely. The quantized coefficients are further encoded using a variation of the Huffman coder described earlier.

Video is a series of images that is presented in order and at a certain rate. Because the images are updated very quickly, the con-

tents of images usually do not change much from one frame to the next. Most video compression algorithms use this temporal correlation between the frames by only transmitting the difference between the current frame and the previous frame after taking into account the motion of objects within a frame.

There are many video compression standards. The most popular video compression standard is known as MPEG2. MPEG stands for Moving Pictures Experts Group, which is the group responsible for this standard. The MPEG2 standard is used in digital cable TV and with DVDs. One of the important properties of the MPEG2 algorithm is its random access capability. This is necessary because the audience may switch the TV at any moment, or may forward or rewind the DVD to any location in the video stream. For random access capability, some frames need to be compressed

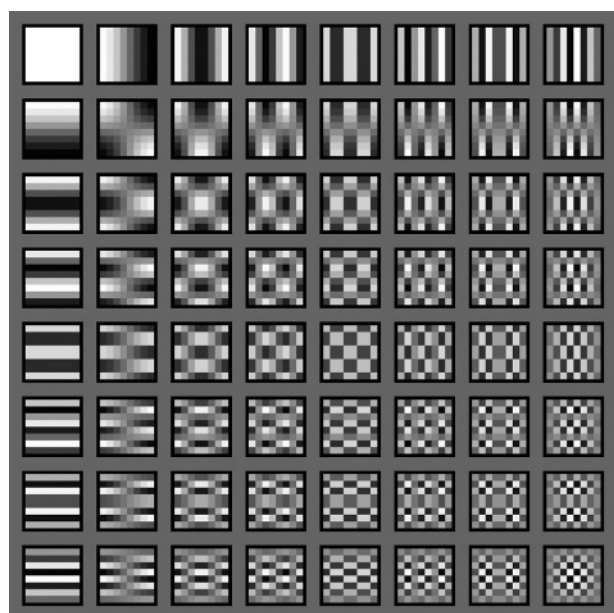


Figure 5 Sixty-Four DCT Basis Functions

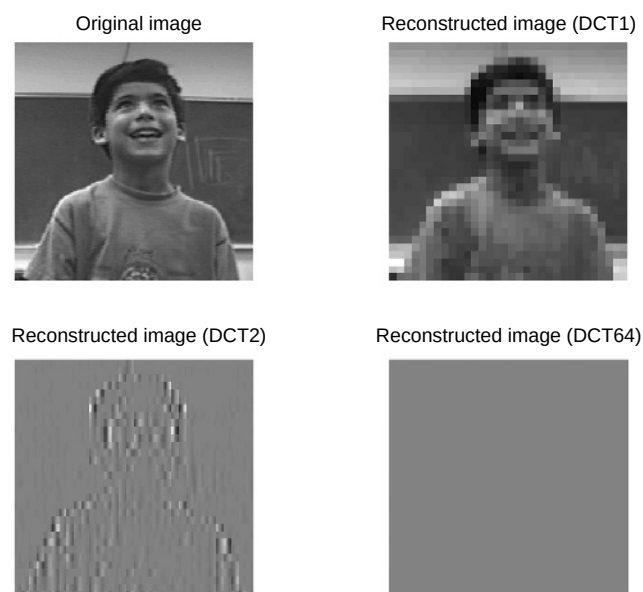


Figure 6 Images Based on DCT Coefficients

periodically without any reference to past frames. These frames are referred to as *I* frames. In order to improve compression efficiency, the MPEG algorithm contains two other kinds of frames, the predictive coded (*P*) frames and the bidirectional predictive coded (*B*) frames. The *P* frames are compressed using motion-compensated prediction from the last *I* or *P* frame, whichever happens to be closest. *I* and *P* frames are called anchor frames.

The *B* frames are coded using motion-compensated prediction from the most recent anchor frames and the closest future anchor frames, as shown in Figure 7. *I* frames achieve the least compression efficiency because they do not use the information from neighboring frames. *B* frames achieve the highest compression efficiency because they use information from the two neighboring frames.

The different frames are organized together to form a group of pictures (GOP), as shown in Figure 7. A GOP is the smallest random access unit in the video sequence. There are many possible structures of a GOP. The GOP structure in Figure 7 is a common one, with 15 frames, and it has the sequence *IBBPBBPBBPBBPBB*. The ratio of *I*, *P*, and *B* pictures in the GOP structure is determined by the nature of the video stream, the bit rate constraints on the output stream, and the required encoding and decoding time.

Because of the reliance of the *B* frame on the future anchor frame, there are two different sequence orders. The display order is the sequence in which the video is displayed to the user as labeled in Figure 7. The other order is the bitstream order in which the frame is processed and transmitted. For the above example, the first frame is an *I* frame, which can be compressed by itself. The next frame to be compressed is the fourth frame, which is compressed based on the prediction from the first frame. Then the second and the third frames are compressed based on the prediction from the first and fourth frames. Hence, the bitstream order of the above example is *IPBBPBBPBBPBB*, and the corresponding display order is 1 4 2 3 7 5 6 10 8 9 13 11 12.

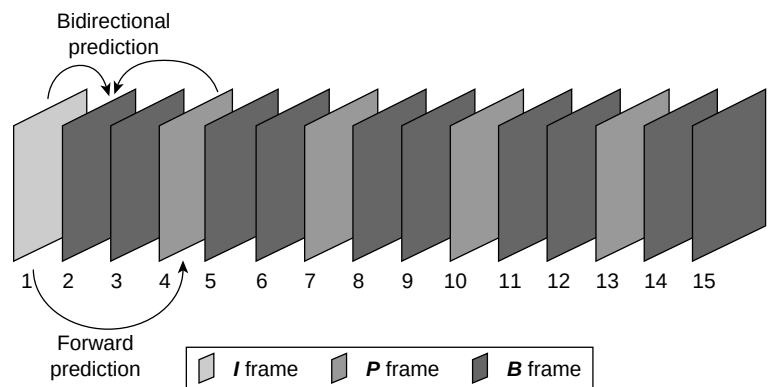


Figure 7 A Possible Arrangement for a Group of Pictures

Summary

This entry presented an overview of compression. We have motivated our discussion by looking at different ways in which structure in a particular source output can be modeled. We have also briefly described some popular compression techniques used for the compression of text, computer files, speech, music, images, and video.

—Dongsheng Bi and Khalid Sayood

See also Data Mining; Variable

Further Reading

- Huffman, D. A. (1952). A method for the construction of minimum redundancy codes. *Proceedings of the IRE*, 40(9), 1098–1101.
- Nelson, M., & Gailly, J.-L. (1996). *The data compression book* (2nd ed.). New York: M&T Books.
- Salomon, D. (2004). *Data compression: The complete reference*. New York: Springer.
- Sayood, K. (2000). *Introduction to data compression* (2nd ed.). San Francisco: Morgan Kaufmann.
- Welch, T. A. (1984, June). A technique for high-performance data compression. *IEEE Computer*, 17(6), 8–19.
- Ziv, J., & Lempel, A. (1977, May). A universal algorithm for data compression. *IEEE Transactions on Information Theory*, 23(3), 227–343.
- Ziv, J., & Lempel, A. (1978, September). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, IT-24(5), 530–536.

Data compression information: <http://datacompression.info>

Data compression principles and practices: <http://www.data-compression.com>

DATA MINING

Data mining refers to the process of discovering useful patterns in very large databases. It uses methods from statistics, machine learning, and database management to restructure and analyze data to extract knowledge or information from the data. Data mining is also known as *knowledge discovery in databases* (KDD).

Data mining is used in a wide range of enterprises. Examples include fraud detection in banking, market segmentation, campaign optimization, genetics research, telecommunications customer turnover (“churn”) prevention, Web site optimization, and crime prevention.

How Data Mining Differs From Traditional Statistical Analysis

Data mining differs from traditional statistical analysis in a number of ways, including amount and type of data used and the goals of the analysis.

Amount of Data

In traditional statistical analysis, the data set being analyzed tends to be small, with tens or hundreds of observations, and rarely more than 10 or 20 variables. For this reason, statistical methods often place a good deal of importance on *statistical efficiency*, or the method’s ability to generate precise estimates with small data sets, in order to minimize the expense of data collection.

In contrast, data mining problems typically use very large databases, with thousands, millions, or even more observations and possibly hundreds or thousands of variables. Because of this, many analysis methods used in data mining focus on computational efficiency, or the ability of the algorithm to process a large number of cases and/or variables in a reasonable amount of time. In some data mining settings, sampling is used to reduce the data set to a manageable size.

Type of Data

In research studies of the type statisticians typically encounter, the data have been specifically planned and carefully collected for the express purpose of the study. The researcher has a good deal of control over the coding of variables, which variables are included in the data, and the format of the resulting data set.

In most data mining settings, the only data available are existing data warehouses or other operational data stores. Formatting and coding of the data have been predetermined by business or organizational needs, usually completely independent of the current analysis. Some fields that the analyst might wish for may be unavailable, and there is usually a lot of irrelevant information in the database that must be sifted through and removed. Because of this, a considerable proportion of effort in a data mining project is devoted to data management, cleaning, and conversion.

Goal of the Analysis

Traditional statistical models are usually used in confirmatory analyses, where the goal of the analysis is to test and confirm or reject specific hypotheses about the process under study. This leads to a strong emphasis on hypothesis testing.

In data mining, however, the goal is usually one of two things: exploratory modeling or practical modeling. *Exploratory modeling* is used to find previously unknown patterns in the data and to generate new hypotheses about the process described by the data. It typically emphasizes models with *transparency*, that is, models that are easy for an analyst to understand and integrate with a theoretical framework.

Practical modeling is used to find ways to optimize a particular process. Here, optimization is loosely defined as improving a business or organizational process in some important way, such as increasing profits for a marketing campaign or adapting a manufacturing process to react more quickly to changes in the economic climate. This kind of modeling is marked by an emphasis on predictive accuracy; for such projects, getting the right answer is generally more important than understanding *how* you got the right answer.

Of course, many data mining projects include both exploration and practical prediction as goals, and so have to balance one against the other to a certain extent.

Origins of Data Mining

Data mining grew largely out of the interaction of the machine learning and database management fields. It wasn't until large databases became widespread that a need was perceived for fast methods of analyzing data and finding patterns. Traditional statistical methods were good at finding structure in data but were not very good at handling the large data sets that enterprises were amassing, so were somewhat limited in their utility. Machine learning researchers developed methods for doing exactly that, and eventually software suites such as SPSS Clementine, IBM Intelligent Miner, and SAS Enterprise Miner were developed to make these machine learning techniques available to data analysts. As computing power became cheap and ubiquitous, and improved mathematical algorithms were discovered, it became more practical to apply some of the more computationally intensive statistical methods to large data sets as well, so they were adopted into the data mining toolbox.

Data Mining Terminology

Because of data mining's diverse ancestry, there is some inconsistency in the terminology to describe various aspects of data mining. Statisticians often have one name for a concept and machine learning theorists a different name, and database modelers a third. Table 1 shows several statistical concepts, along with other names by which the concepts are known to data miners.

The Data Mining Process

As indicated above, data mining is a complex process, of which model building is only a part. The process

Table 1 Data Mining Terminology

<i>Statistical Concept</i>	<i>Synonymous Data Mining Terms</i>
Variable	Field, Attribute, Feature, Column, Dimension
Case	Record, Row, Example, Instance
Dependent variable	Target (field), Output (field), Outcome (field)
Independent variable	Predictor (field), Input (field), Attribute
Nominal	Categorical, Set, Discrete
Continuous	Numeric, Real
Model estimation	Model building, Training, Teaching, Learning
Predictive modeling	Supervised learning
Clustering, exploratory factor analysis, other nonpredictive model building	Unsupervised learning
Outlier	Anomaly
Loss	Cost, Utility

includes understanding the problem and the data to be mined, preparing the data for mining, building models, evaluating the results of modeling, and deploying final models to directly address the original problem. The Cross Industry Standard Process for Data Mining (CRISP-DM) consortium has developed a detailed model for the complete data mining process. You can find out more about the model in the Further Reading at the end of this entry.

Modeling Procedures Used in Data Mining

Data miners use a wide variety of modeling methods borrowed from machine learning and statistics, as well as methods developed specifically for data mining. Some of the more popular methods are briefly described as follows.

Decision Trees

This family of algorithms uses recursive partitioning to split the data into subgroups, based on values of the predictor variables, such that each subgroup tends to have similar values for the target variable. The effect of the recursive partitioning is to build a "tree" of nodes, each representing a subgroup. The various methods differ primarily in the loss function used to

decide each split, typically a statistical, entropy, or information criterion. Examples include CHAID, CART, C5.0, and QUEST.

Rule Induction

This family of algorithms derives rules of the form *IF* x *THEN* y from the data, where x is a set of *antecedent* conditions (values for predictor fields) and y is a *consequent* condition [value(s) for the target field(s)]. Such a rule can also be written as $x \rightarrow y$. For example, in a market basket analysis analyzing items purchased together at a store, you may get rules of the form *BREAD* $\rightarrow$ *BUTTER* (0.60), meaning that 60% of the customers who buy bread also buy butter. In this example, 60% is called the *confidence* for the rule *BREAD* $\rightarrow$ *BUTTER*. Examples of rule induction algorithms include Apriori and Carma.

Neural Networks

This family of algorithms attempts to simulate, in an abstract fashion, the way nerve cells process information in the brain. A neural network model consists of a set of interconnected nodes, each of which processes information in a simple way and then passes it on to other nodes to which it is connected, something like a (simplified) biological neuron. A neural network learns by adjusting the strengths of the connections between nodes (the weights), either to improve predictive performance or based on an internal loss function. The most common types of neural networks are backpropagation networks (also called *multilayer perceptrons*), radial basis function networks, and Kohonen networks (also called *self-organizing maps*).

Genetic Algorithms

Genetic algorithms work using evolutionary principles to select good models from a population of candidate models. An initial set of models is generated, usually randomly, and each model is tested for accuracy and/or other goodness criteria. Some number of models, those that give the best performance, are

saved, and the rest are discarded or *culled*. New models are created by randomly combining characteristics of the models saved from the previous round, analogous to sexual reproduction, and the models are evaluated again. In some cases, *mutations* are also added to each successive generation, where model parameters are changed randomly, independent of any other observed values in the saved candidate models from the previous iteration. The process then repeats, with the candidate models being culled, recombined, and tested again. This process, analogous to natural selection, will tend to produce better and better models with each successive generation. Iterations stop when one or more of the candidate models satisfies some criterion for judging the model(s) as adequate.

Note that genetic algorithms are a basic optimization method and can be used with a variety of model types.

Support Vector Machines

A support vector machine (SVM) is a method of constructing models based on kernel methods. An SVM works by projecting the input space (the space defined by the input fields or dimensions) into a very high-dimensional feature space, in which a linear discriminator or hyperplane (or a set of hyperplanes) can be used to separate target subgroups from one another. The use of kernel functions saves SVMs from having to represent the full high-dimensional feature space. SVMs use the maximum margin hyperplane in the feature space, which generally controls overfitting, in spite of the high dimensionality of the feature space.

Various kernels have been used with SVMs, including radial basis function, linear, polynomial, and Gaussian.

Nearest Neighbor Methods

This family of methods, also known as memory-based reasoning, uses examples from the training data to classify or make predictions for new data. A new record is compared to a table of exemplars from the training data, and the target value of the stored record

most similar to the new record is used as the predicted value (or classification). In some cases, the k most similar records are identified, and the prediction is some combination of the target values for those records.

Clustering

Clustering methods attempt to identify homogeneous groups in the data set, usually without respect to known categories or any particular target field. Because there is no known correct answer, clustering is often referred to as unsupervised learning. (Predictive models are supervised because they usually involve comparing a predicted value to an observed target value and adjusting the model to reduce the distance between the two.) Clustering is often used to define group membership, which is then used to predict some other characteristic in a subsequent model. It can also be used to detect outliers or anomalies by highlighting records that don't fit well with any identifiable subgroup. The two most common types of clustering are *hierarchical clustering* and *k-means clustering*.

Hierarchical clustering works by starting with each record defining a separate cluster. The two most similar clusters are merged into a compound cluster. Then, the next two most similar clusters are merged, and so on until all records have been merged into one giant cluster. By selecting a cut-point in minimum intercluster distance, you can define a cluster solution with any number of clusters you want.

K-means clustering starts by defining the number of clusters a priori. A set of k randomly selected records defines the initial cluster centers. All of the other records in the data set are assigned to one of the k clusters: The assigned cluster is the closest cluster based on the Euclidean distance between the record and the cluster center. After the records have been assigned, a new cluster center is calculated for each cluster as the mean of the input fields for all records assigned to the cluster. Records are then reassigned to their closest clusters, and cluster centers are recalculated. Iterations continue until the clusters stabilize, that is, until no records need to be reassigned to a different cluster after cluster centers are updated.

Text Mining

Text mining refers to methods that attempt to find relationships and patterns in unstructured text, such as newspaper articles, Web sites, and academic papers. Text mining usually includes categorization of words or phrases and identification of related words and concepts. It is a fairly low-level kind of analysis and does not purport to automatically "understand" the texts being analyzed. However, by modeling relationships between word categories, often new information or hypotheses can be generated.

Statistical Methods

Several statistical methods have become widely used in data mining, including linear regression, logistic regression, and discriminant analysis.

—Clay Helberg

See also Artificial Neural Network; Cluster Analysis; Discriminant Analysis; Linear Regression; Logistic Regression Analysis; Support Vector Machines

Further Reading

- Berry, M., & Linoff, G. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed.). New York: Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
- Pyle, D. (1999). *Data preparation for data mining*. San Francisco: Morgan Kaufmann.
- Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

The CRISP-DM consortium: <http://www.crisp-dm.org/>

The Data Mine, a data mining-oriented Web site: <http://www.the-data-mine.com/>

KD Nuggets, a clearinghouse for data mining information: <http://www.kdnuggets.com/>

Kurt Thearling's data mining page: <http://www.thearling.com/>

Two Crows: <http://www.twocrows.com/>

DECISION BOUNDARY

A decision boundary is a partition in n -dimensional space that divides the space into two or more response

regions. A decision boundary can take any functional form, but it is often useful to derive the optimal decision boundary that maximizes long-run accuracy.

The use of decision boundaries is widespread and forms the basis of a branch of statistics known as discriminant analysis. Usually, discriminant analysis assumes a *linear* decision bound and has been applied in many settings. For example, the clinical psychiatrist might be interested in identifying the set of factors that best predicts whether an individual is likely to evidence some clinical disorder. To achieve this goal, the researcher identifies a set of predictor variables taken at Time 1 (e.g., symptoms, neuropsychological test scores, etc.) and then constructs a linear function of these predictors that best separates depressed from nondepressed or schizophrenic from nonschizophrenic patients diagnosed at Time 2. The resulting decision bound then can be applied to symptom and neuropsychological test data collected on new patients to determine whether they are at risk for that clinical disorder later in life. Similar applications can be found in machine learning (e.g., automated speech recognition) and several other domains.

To make this definition more rigorous, suppose we have two categories of clinical disorders, such as depressed and nondepressed individuals with predictor variables in n -dimensional space. Denote the two multivariate probability density functions $f_D(x)$ and $f_{ND}(x)$ and the two diagnoses R_D and R_{ND} . To maximize accuracy, it is optimal to use the following decision rule:

$$\text{If } f_D(x)/f_{ND}(x) > 1, \text{ then } R_D, \text{ else } R_{ND}. \quad (1)$$

Notice that the optimal decision bound is the set of points that satisfies

$$f_D(x)/f_{ND}(x) = 1.$$

It is common to assume that $f_D(x)$ and $f_{ND}(x)$ are multivariate normal. Suppose that μ_D and μ_{ND} denote the depressed and nondepressed means, respectively, and that Σ_D and Σ_{ND} denote the multivariate normal covariance matrices. In addition, suppose that $\Sigma_D = \Sigma_{ND} = \Sigma$. Under the latter condition, the optimal decision bound is linear.

Expanding Equation 1 yields

$$\begin{aligned} f_D(x)/f_{ND}(x) &= 1 \\ &= \frac{(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \mu_D)' \Sigma^{-1}(x - \mu_D)\right]}{(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \mu_{ND})' \Sigma^{-1}(x - \mu_{ND})\right]} \quad (2) \\ &= \exp\left[-\frac{1}{2}(x - \mu_D)' \Sigma^{-1}(x - \mu_D) + \frac{1}{2}(x - \mu_{ND})' \Sigma^{-1}(x - \mu_{ND})\right] \end{aligned}$$

Taking the natural log of both sides of Equation 2 yields

$$\begin{aligned} h(x) = \ln [f_D(x)/f_{ND}(x)] &= (\mu_{ND} - \mu_D)' \Sigma^{-1}x \\ &+ \frac{1}{2} (\mu_D' \Sigma^{-1} \mu_D - \mu_{ND}' \Sigma^{-1} \mu_{ND}), \end{aligned} \quad (3)$$

which is linear in x .

As a concrete example, suppose that the objects are two-dimensional with $\mu_D = [100 \ 200]'$, $\mu_{ND} = [200 \ 100]$, $\Sigma_D = \Sigma_{ND} = \Sigma \ 50I$ (where I is the identity matrix). Applying Equation 3 yields

$$.04x_1 - .04x_2 = 0.$$

—W. Todd Maddox

See also Discriminant Analysis; Discriminant Correspondence Analysis

Further Reading

- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Fukunaga, K. (1972). *Introduction to statistical pattern recognition*. New York: Academic Press.
- Morrison, D. F. (1967). *Multivariate statistical methods*. New York: McGraw-Hill.

DECISION THEORY

Every day, a multitude of decisions are made that affect not only a small number of individuals, but also potentially millions of people. These decisions, which take place in hospitals, pharmaceutical companies, government offices, investing companies, and so on,

are made based on incomplete information and under various conditions of uncertainty as to the ability of the decision maker(s) to follow through with the commitments made. Thus, nations declare wars with incomplete information as to the capabilities of the others involved in the war, and with uncertainty as to what the impact of these actions might be on their own economies and citizens. A pharmaceutical company must decide whether to market a new drug based on limited information resulting from the clinical trials and economic uncertainty as to whether the drug can compete in the market with other drugs.

Decision theory consists of techniques, ideas, and methodologies that are appropriate for helping the decision maker to reach a decision in an optimal fashion in the face of uncertainty. Given the universality of decision theory in corporate life, government action, and everyday life, it is not surprising to find that decision theory has been embraced by almost every scientific discipline. Thus, game theory permeates the theory and applications in economics. Psychologists know game theory as the theory of social interactions, and political scientists study rational choice theory.

All of these approaches to decision making have several essential elements in common that will be discussed below in the context of statistical decision theory. Game theory served as the precursor to most of the ideas in decision theory, and its place in modern decision making was cemented by John von Neumann and Oskar Morgenstern's fundamental work on the *Theory of Games and Economic Behavior* (1944). Although decision theory developed from game theory, there is a fundamental difference between the two. Informally, whereas in game theory, players make decisions based on their beliefs of what other players—whose interests may be diametrically opposed to theirs—will do, decision theory concerns itself with the study of decisions of individuals unconcerned with the plans of others—their “opponent” being nature.

Wald unified at once ideas from game theory and Neyman's and Pearson's mathematical developments in the theory of statistics in his elegant work *Statistical Decision Functions*. It is this approach on which the rest of the discussion focuses.

Statistical Decision Theory

There are at least three common elements to all introductory courses of statistical theory and methodology: estimation, hypothesis testing, and confidence intervals. As taught in an introductory course, these three topics may appear as being unrelated. Toward the end of the course, however, the student learns to “invert” acceptance regions to obtain confidence intervals, and also learns to use confidence intervals to carry out tests of hypothesis. In addition, confidence intervals are introduced as point estimates together with a measure of precision of the estimates, typically 2 or 3 standard errors of the point estimate. Statistical decision theory unifies these ideas, and others, into one paradigm.

Let $X_1, \dots, X_n$ represent the data observed as the outcome of an experiment E and let F represent the distribution of $(X_1, \dots, X_n)$, which we assume to be parametrized by θ , where θ may be a vector of parameters, and the set of all possible values of θ , called the *parameter space*, is denoted by Θ . This dependence of F on θ will be denoted as F_θ . The objective is to use $(X_1, \dots, X_n)$ to make inferences about θ . Faced with this problem, the statistician considers all the possible actions A (A is called the *action space*) that can be taken and makes a decision based on a criterion that involves minimizing the expected loss. This requires the statistician to define a function, the *loss function*, which represents the loss when the true state of nature is θ and the statistician decides for action a . This *loss function* is denoted as $L(\theta, a)$, and L is usually selected to be of the form

$$L(\theta, a) = v(a - \theta),$$

$$v(0) = 0,$$

$$v(x) \text{ is increasing in } |x|.$$

Examples of loss functions include the following:

$$\text{Squared error loss: } L(\theta, a) = (a - \theta)^2,$$

$$\text{Absolute error loss: } L(\theta, a) = |a - \theta|, \text{ and}$$

$$\text{Linex error loss: } L(\theta, a) = b(e^{c(a - \theta)} - c(a - \theta) - 1).$$

Historically, squared error loss has been used mostly because of the relative ease with which properties of the resulting procedures can be analyzed. It is

common to assume symmetry of the loss function. This implies implicitly that overestimating by an amount d has the same consequences as underestimating by the same amount. Varian, in the context of real estate assessment, argued the need for the Linex loss as more representative of losses accrued when assessing the value of real estate.

Once the probability model for $(X_1, \dots, X_n)$ and the action space A have been defined, and the loss function has been selected, the statistician attempts to make a decision $a(X_1, \dots, X_n)$ that minimizes the risk function

$$R(\theta, a(X_1, \dots, X_n)) = E_\theta(L(\theta, a(X_1, \dots, X_n))),$$

for all θ and for all $a^* \neq a$. Of course, this is not possible because if $\theta^* \in \Theta$, then no other $a(X_1, \dots, X_n)$ can be better than the action $a^*(X_1, \dots, X_n) = \theta^*$, which disregards the data and always “guesses” θ^* . Thus, a need arises for eliminating procedures that “pay too much attention” to certain models and completely disregard others. Moreover, even when these “partial” estimators are eliminated, it usually happens that risk functions for different decision rules will, as functions of θ , crisscross, and then it is not clear which decision to use. Before discussing other ideas in decision theory, let us consider the following example.

Example

Consider the normal model with mean $\theta \in \mathbb{R}$, and known variance s^2 , and consider the three following problems:

1. Point estimation of θ : Here, $\Theta=A=\mathbb{R}$. Choosing squared error as the loss function, the risk function is given by

$$R(\theta, a(X_1, \dots, X_n)) = \text{Variance}(a(X_1, \dots, X_n)) + \text{Bias}(a(X_1, \dots, X_n))^2,$$

where $\text{Bias}(a(X_1, \dots, X_n)) = E_\theta(a(X_1, \dots, X_n)) - \theta$. This risk function is also known as the *mean squared error* of the estimator. In this case, for example, the

sample mean $a(X_1, \dots, X_n) = \bar{X}_n$, being *unbiased*, has a risk function equal to $\frac{\sigma^2}{n}$.

2. Hypothesis testing: Consider testing the simple hypothesis $H_0 : \theta = \theta_0$ versus the simple hypothesis $H_1 : \theta = \theta_1$. In this setup, there are only two possible actions:

$$A = \{\text{Do not reject } H_0, \text{Reject } H_0\}$$

Let a *loss function* be defined as follows:

$$L(\theta, \text{Do not reject } H_0) = 0 \text{ when } \theta = \theta_0,$$

$$L(\theta, \text{Do not reject } H_0) = k_0 \text{ when } \theta = \theta_1,$$

$$L(\theta, \text{Reject } H_0) = k_1 \text{ when } \theta = \theta_0,$$

$$L(\theta, \text{Reject } H_0) = 0 \text{ when } \theta = \theta_1.$$

Note that the loss function is zero when the correct decision is made, the loss is k_0 if H_0 is not rejected when it should be rejected (an error of Type II), and the loss is k_1 if H_0 is rejected when it should not be rejected (an error of Type I), where the decisions of Rejecting H_0 and Not rejecting H_0 are defined in terms of the value of $a(X_1, \dots, X_n) = \bar{X}_n$. The problem of selecting the test function based on $a(X_1, \dots, X_n)$, that minimizes $P\{\text{Type II error}\} = P_{\theta_1}\{\text{Do not Reject } H_0\}$, subject to the condition that $P\{\text{Type I error}\} = P_{\theta_0}\{\text{Reject } H_0\} \leq \alpha$, for some preselected α , was addressed by Neyman and Pearson, and the solution is their fundamental Neyman-Pearson lemma (1938).

3. Confidence intervals for θ : Let A consist of all intervals $(\underline{a}(X_1, \dots, X_n), \bar{a}(X_1, \dots, X_n))$ with $\underline{a} < \bar{a}$. Let the loss function be of the form

$$L(\theta; \underline{a}, \bar{a}) = L_1(\theta, \underline{a}) + L_2(\theta, \bar{a})$$

where L_1 is nonincreasing in $\underline{a}$ for $\underline{a} < \theta$ and 0 for $\underline{a} \geq \theta$ and L_2 is nondecreasing in $\bar{a}$ for $\bar{a} > \theta$ and 0 for $\bar{a} \leq \theta$. The goal is then to find the interval $(\underline{a}^*, \bar{a}^*)$ that minimizes the risk function $E_\theta(L(\theta; \underline{a}, \bar{a}))$ subject to

$$P_\theta\{\underline{a} > \theta\} \leq \alpha_1 \text{ and } P_\theta\{\bar{a} < \theta\} \leq \alpha_2.$$

One example of a loss function of the type $L_1 + L_2$ is the function that takes the length $\bar{a} - \underline{a}$ of the interval as the loss. Taking, for example, $a_1 = a_2 = .025$, it is well-known that the usual 95% confidence interval

$$\left(\bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

for θ in the normal case is optimal in the sense of minimizing the expected length subject to the constraint that

$$P_\theta\{\underline{a} > \theta\} \leq .025 \text{ and } P_\theta\{\bar{a} < \theta\} \leq .025.$$

In what follows, we will restrict attention to the case of point estimation, although most remarks to be made also apply more generally. As previously discussed, once the elements (action space, probability model, and loss functions) of the statistical decision theoretic problem have been selected, the statistician would like to choose a decision rule $a(X_1, \dots, X_n)$ that, uniformly, in θ and in $a^* \in A$, minimizes the risk $E_\theta(L(\theta, a))$. That is, the goal is to find the decision rule $a(X_1, \dots, X_n)$ such that

$$E_\theta(L(\theta, a(X_1, \dots, X_n))) \leq E_\theta(L(\theta, a^*(X_1, \dots, X_n))).$$

for all $\theta \in \Theta$ and all $a^* \in A$. There is, however, the difficulty alluded to earlier. It is not possible to carry out this program because there are decision rules that pay too much attention to some values of θ while disregarding most other values of θ . At this juncture, there are two possible ways to proceed. First, one may restrict the class of decision rules under consideration, for example, by eliminating those that only pay attention to a few points in parameter space. Thus, for example, restricting attention to decision rules that are unbiased eliminates many estimators, and the hope is that in this smaller class of decision rules, one may find one that is uniformly best. There is an extensive literature on finding the uniformly minimum variance unbiased estimators (UMVUE). When the problem satisfies certain symmetry properties, another way of restricting the class of estimators to a more

manageable size with the hope of finding an estimator that is uniformly best in the reduced class is to consider only those estimators that are *equivariant*.

The second approach defines preference orders for the risk functions. One possible way of doing this is to integrate $R(\theta, a(X_1, \dots, X_n))$ with respect to a probability distribution $\Pi(\theta)$ on Θ . This approach gives rise to Bayes estimators, and more generally to the Bayesian approach that interprets the distribution $\Pi(\theta)$ as representing the statistician's "prior" knowledge about Θ . Thus, informally, a Bayes estimator for θ with respect to a loss function $L(\theta, a)$ achieves the smallest area under the weighted risk function where the weight is provided by the specific prior distribution Π on Θ . That is, a Bayes estimator $\delta_\pi(x)$, with respect to the prior distribution $\Pi(\theta)$ and the loss function $L(\theta, a)$, minimizes the Bayes risk and therefore,

$$\int_{\Theta} R(\theta, \delta_\pi(x)) d\Pi(\theta) = \inf_{\delta \in A} \int_{\Theta} R(\theta, \delta(x)) d\Pi(\theta).$$

Alternatively, one may order risk functions, and hence decision rules, by preferring decision rule a_1 to decision rule a_2 if

$$\sup_{\theta \in \Theta} R(\theta, a_1) \leq \sup_{\theta \in \Theta} R(\theta, a_2).$$

A decision rule $a^* \in A$ is then said to be *minimax* if

$$\sup_{\theta \in \Theta} R(\theta, a^*) = \inf_{a \in A} \left\{ \sup_{\theta \in \Theta} R(\theta, a) \right\}.$$

Thus, a minimax estimator minimizes, among all estimators, the maximum risk.

Let the loss function be squared error. Consider the estimator $\bar{X}_n$. Its risk is constant and given by $\frac{\sigma^2}{n}$. Because $\bar{X}_n$ is unbiased, it is of interest to determine if it is also best in the class of unbiased estimators. That is, is it UMVUE? It turns out that the estimator is UMVUE, and also minimax. However, it is not Bayes with respect to any prior distribution on Θ . In fact, except for very few cases, an unbiased estimator cannot be Bayes.

However, $\bar{X}_n$ is the limit of a sequence of Bayes estimators. More precisely, consider as a prior distribution on Θ the normal distribution with mean m and variance d^2 . That is, the prior density is given as follows:

$$\lambda(\theta) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2d^2} (\theta - \mu)^2 \right\}.$$

The Bayes estimator with respect to this prior and squared error loss is

$$a(X_1, \dots, X_n) = \left\{ \frac{(n/\sigma^2)}{n/\sigma^2 + 1/d^2} \right\} \bar{X}_n + \left\{ \frac{(1/d^2)}{n/\sigma^2 + 1/d^2} \right\} \mu.$$

Thus, the Bayes estimator is a convex combination of the sample mean $\bar{X}_n$ and m , the mean of the prior distribution. Letting $b \rightarrow \infty$, it is seen that $\bar{X}_n$ arises as a limit of Bayes estimators.

—Javier Rojo

See also Bayesian Statistics; Evidence-Based Practice

Further Reading

- Bergus, G. R., & Hamm, R. (1995). Clinical practice: How physicians make medical decisions and why medical decision making can help. *Primary Care*, 22(2), 167–180.
- Braddock, C. H., III, Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *Journal of the American Medical Association*, 282(24), 2313–2320.
- Green, D., & Shapiro, I. (1994). *Pathologies of rational choice theory: A critique of applications in political science*. New Haven, CT: Yale University Press.
- Lehmann, E. L. (1983). *Theory of point estimation*. New York: Wiley.
- Lehmann, E. L. (1991). *Testing statistical hypotheses*. Belmont, CA: Brooks/Cole.
- Lehmann, E. L. (2004). Optimality and symposia: Some history. In J. Rojo & V. Perez-Abreu (Eds.), *The First Erich L. Lehmann Symposium: Optimality*. IMS Lecture Notes and Monograph Series, Vol. 44.
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses.

Philosophical Transactions of the Royal Society, Series A, 231, 289–337.

- Rojo, J. (1987). On the admissibility of $c\bar{X} + d$ with respect to the linex loss function. *Commun. Statist. Theory and Meth.*, 16, 3745–3748.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, 28, 1–39.
- Varian, H. R. (1975). A Bayesian approach to real estate assessment. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* (pp. 195–208). Amsterdam: Elsevier North-Holland.
- Wald, A. (1950). *Statistical decision functions*. New York: Wiley.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81, 446–451.

DELPHI TECHNIQUE

The Delphi technique is a means of collecting data from a diverse group of people for the purpose of reaching a consensus. This entry presents the basic process of the Delphi technique as well as some variations on the process that can be used to meet specific needs. Although the Delphi technique allows for refinement of original ideas and therefore promotes high-quality decisions, it can be time consuming and subject to bias. Examples of how the process can be used in research are provided. From this entry, readers can determine if the Delphi technique is appropriate for their particular situations.

According to S. J. Adams, the Delphi technique provides a representation of varied backgrounds, and it prevents individuals with strong personalities from dominating a group. The purpose is to obtain information from participants to help in the areas of problem solving, planning, and decision making. The Delphi technique is a way to reach a consensus among a group of experts.

The RAND Corporation developed the technique during the 1950s as an approach to forecasting the likelihood and the potential impact of Russian bombing attacks on the United States. The approach was named for the Oracle of Delphi of Greek mythology. It was soon adopted by technological forecasting experts and eventually found its way into other types of research.

Theoretical Basis for the Delphi Technique

Consensus Theory

According to a variety of researchers, the objective of users of the Delphi technique is to achieve consensus. Some proponents of consensus theory believe that building consensus offers opportunity for communal renewal and for achieving group commitment to common goals.

Anonymity

Some researchers and theorists believe that anonymity is helpful for generating quality ideas. Others expect that using the Delphi process discourages individual dominance and simultaneously encourages each person to share his or her ideas without fear of intimidation.

Divergent Thought

Some researchers observe that divergent thinking occurs when individuals or groups are introduced to minority opinions. Anonymity and exposure to a variety of viewpoints contributes to improved creativity and decision making.

Purpose and Uses in Research

Researchers have used the Delphi technique for gathering broad-based opinions from experts, refining their views, and reaching consensus on predictions and plans for dealing with complex issues. The data generated have been used in forecasting, public budgeting, and goal setting. Decision makers in such diverse disciplines as education, safety management, family therapy research, environmental studies, government, medicine, and community health have relied on Delphi for all or portions of their research data.

Delphi Technique Process

Delphi technique involved several carefully structured steps. It bore some resemblance to Nominal Group Technique (NGT) in that with both processes,

individual contributions were made anonymously. However, the standard format for Delphi did not require participants to meet. Thus, not only were responses anonymous, but even the identity of other participants might be unknown to the group. The procedure involved two to four rounds of responses. However, prior to the first round, primary stakeholders had to do the following:

1. *Select a monitor or monitor group.* This person or persons should be experts both on the topic and on written communication skills.
2. *Select participants.* Participants usually were stakeholders as well. However, they could be noninvolved experts.
3. *Invite participants.* Selected participants were invited by telephone, mail, or e-mail to take part in the process.
4. *Develop a broad question or statement for consideration.* The monitor developed the initial question or statement, perhaps in conjunction with other stakeholders.

To begin the rounds, the monitor was responsible for (a) identifying and orienting participants; (b) getting the question to each participant; (c) receiving input from each participant; (d) summarizing the information; (e) sending the summary and a new, more focused question to the participants; and (f) determining that no more rounds were needed. The process concluded with a resolution. When consensus was reached, the resolution was announced to participants. Panel participants committed to the decision (see Figure 1).

It was recommended that 12 to 15 panel members were an appropriate size. Panel sizes ranged from a few to hundreds of members, depending on the research topic. A response rate of 70% or greater was typically acceptable. It was common for the iteration process to last only two or three rounds before consensus was reached.

Statistical Measures of Agreement

Panels commonly have used Likert scales to assess the rating of items. The Delphi monitor calculated

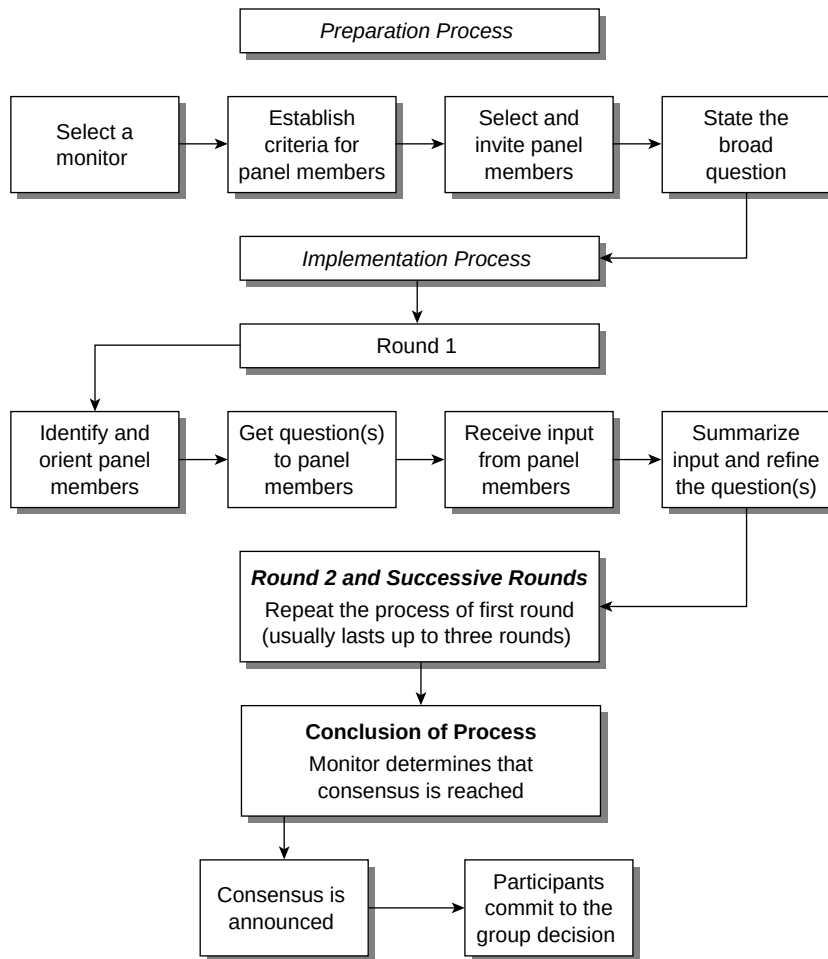


Figure 1 Flowchart of Delphi Technique

summary statistics, such as the median, following each round and reported them to the panel members for consideration during the next round. Researchers found the median to represent the most common value provided by a panel member and cited the interquartile range (the middle half of the scores) as a measure of consensus. The smaller the interquartile range, the greater the consensus. Several studies supported using the median and the interquartile range as measures of agreement and consensus.

Delphi scholars reported means to panel members in successive rounds and standard deviations as measures of consensus. However, other researchers suggested that providing standard deviations to expert panels was misleading because they were not applicable to ordinal data.

Variations of the Delphi Technique

Real-Time and Policy Delphi

Clayton identified three types of Delphi: (a) conventional, (b) real time, and (c) policy. The conventional process was presented earlier. Real-time Delphi differed in that it occurred face-to-face, within the context of a meeting or conference. Policy Delphi asked participants for information on which a decision was to be made. Policy Delphi and real-time Delphi could be combined. Consensus was not an objective in either case.

Combinations of Delphi With Nominal Group Technique

The possibility of voting when consensus could not be reached could be considered a variation on Delphi. It was a compromise of the ideal Delphi and borrowed the last phase from NGT.

Mixed Methods Studies

Some researchers have used mixed methodologies to complement one another. For example, one might use Delphi to determine items for a questionnaire that would be used in a mail survey regarding environmental policy.

Historic Delphi

In this variation, Strauss and Zeigler found that participants attempted to apply systematically the writings of classical political philosophers to current issues. That interesting study is presented in the *Historic Delphi Approach* section of this entry.

Reactive Delphi

A popular variation of Delphi was the reactive method, in which panel members reacted to pregenerated items or questions in Round 1, rather than producing a list of ideas. In this variation, the researcher prepared a list of items from a review of related literature, and the subject matter experts were asked to rate the importance of each item on some scale. The researchers found that such structured first rounds diminished the assessment properties found in the classic Delphi technique.

Advantages and Disadvantages of the Delphi Technique

Advantages of Delphi Technique

Effective structure. The process allowed participants to refine their original ideas. That resulted in high-quality decisions on complex issues. These results came from professionals who gained insights from one another's input during the successive rounds.

Fiscal economy. Little financial cost was involved in using the conventional Delphi technique. There are no travel costs, no need for accommodations, and usually no stipends for participants.

Collaboration. Several researchers noted that in reaching consensus, the Delphi technique fostered collaboration among parties who would be needed to carry out the group's decisions.

Disadvantages and Limitations of the Delphi Technique

Time requirements. Two aspects of time apply to Delphi. First, there is the requirement of the monitor(s) and of each participant. Also, the transmission of ideas could result in an overall time frame of several weeks. Difficulty in retaining participants throughout the process may be a problem. Sometimes, the latter could be lessened by using e-mail. However, that would require special care to maintain anonymity.

Inadequacy as sole method. According to several studies, when used alone, the Delphi technique was inadequate for forecasting. Also in this regard, one

must consider the built-in dangers of bias. A discussion of such dangers follows.

Central tendency. Michigan State University Extension found that consensus building generally has involved finding a middle-of-the-road position, eliminating the extreme ends of the spectrum. This feature has caused some groups to feel that their views were rejected and that the process was rigged.

Bias. It is important to ensure that experts are not influenced by the researcher's objective. To guard against this, Delphi experiments usually use two or more separate groups of experts. There are numerous opportunities for introducing bias into the Delphi process, including (a) setting eligibility standards and soliciting participants, (b) formulating the questions, (c) summarizing participants' contributions, (d) rephrasing questions for successive rounds, (e) determining the number of rounds, (f) phrasing consensus statements, and (g) fostering commitment to decisions. For each step in the process, moderators must be accountable for keeping the process bias free.

Communication difficulties. Strauss and Zeigler discussed the possibility of misunderstandings, noting that the respondents may misunderstand the brief written inputs of the panelists. Others noted that Delphi relied heavily on the written communication skills of experts. This made the selection of participants especially difficult, because expertise in the field did not necessarily include communication expertise.

Ethical standards and need for trust. Conflict could arise from the many opportunities for bias using Delphi. Root causes could be intended or unintended bias or lack of group trust in the process. Without mutual trust, it would be impossible to reach consensus. Three types of disagreement involving ethics and trust are especially noteworthy:

1. *Forecast versus foresight.* Forecasting with the Delphi technique is used to predict what is likely to occur, whereas foresight implies that the process is guided toward a predetermined outcome. As mentioned elsewhere, the Delphi process cuts off extreme views and seeks a middle ground. Unfortunately, those whose opinions have been sacrificed may feel disenfranchised. Thus, when a facilitator records a

group's comments, the final outcome can be highly questionable.

2. *Consensus versus coalition.* The same principle applies here as to the forecasting versus foresight outcomes.

3. *Consensus and morality.* Among populations that must interrelate, pluralism was the only viable option. However, if Delphi technique participants represented a full spectrum of ethical values, absolutism, pluralism, and relativism, some participants might be open, some might be reluctantly persuaded, and some might feel excluded. If so, consensus could not be reached. However, the Delphi technique was designed especially to reach consensus on complex issues, and complex questions almost always involve moral values.

In a discussion of moral consensus, the following questions should be considered: Does consensus carry any moral authority? and Can groups ever achieve a valid consensus on issues of bioethics? Their considerations have raised questions about both the practicality and the propriety of using the Delphi technique to address the very issues for which the technique was designed. Perhaps one should consider the words of Mohandas Gandhi, "In matters of conscience, the law of majority has no place."

Research Applications

In this section is a brief overview of some of the institutional research applications using Delphi. They range from employee issues, such as selecting job candidates and handling occupational stress, to forecasting training needs and needed changes in human resource practices, to needs assessment.

Frazer and Sechrist examined the effects of occupational stress on employees in nuclear medicine, radiologic technology, and medical technology. They used the Delphi technique to determine 35 job stressors for each discipline. Improved communication strategies and managerial development were noted as solutions to occupational stress.

Olmstead-Schafer, Story, and Haughton used the Delphi method to forecast training needs of public

health nutritionists. It was the consensus of their panel that communication, policy development, and managerial skills be included in the curriculum for training nutrition professionals.

Japanese firms used the Delphi method in forecasting needed changes in human resource practices. The panel made predictions regarding the year in which strongly held Japanese institutions of lifelong employment, seniority-focused compensation, and promotion from within at the exclusion of external recruiting would transition to practices consistent with Western cultures. The overall consensus was that it would take two decades to see significant changes.

Tavana, Kennedy, and Joglekar studied the effectiveness of the Delphi approach for ranking job candidates for a nursing management position. After two rounds, the experts reached consensus on the top applicant from a field of seven. Schuler found that the Delphi approach was beneficial in emergent and less structured subject areas such as human resource planning.

Finally, program evaluation was noted as another area for using Delphi. The Delphi technique is particularly useful for studies requiring a needs assessment.

Historic Delphi Application

Strauss and Zeigler conducted an interesting historic Delphi study. Their objective was to systematically scrutinize the great political philosophers of the past and to apply their wisdom to contemporary problems. Plato, Aristotle, Hobbes, Machiavelli, Swift, Burke, Rousseau, Locke, Marx, and Freud were the philosophers. Ten panels of six experts each (mostly university professors) represented the philosophers. The questionnaire contained 42 problem statements regarding serious issues in Western society, and each statement had a three-part question:

1. In general, what was political philosopher X's view on problem statement Y?
2. Based on your knowledge of political philosopher X, how would he have reacted to the problem statement in his own time?
3. If alive today, how would political philosopher X resolve the problem?

The second round consisted of multiple-choice options. The experts responded on a 5-point Likert scale that asked to what extent they agreed or disagreed with each item. The items in the second round were taken from each group's first-round responses.

The product of this academic exercise was a series of options for handling a variety of social problems based on Western philosophical thought. Strauss and Zeigler hoped that, in addition to accomplishing this pragmatic objective, their development of the historic Delphi approach would be a meaningful way for students to study philosophy.

Comparison of Delphi, Nominal Group, and Q-Sort Techniques

Delphi and NGT have many similarities. Each encourages divergent thought, preserves anonymity of participants' contributions, and is aimed at consensus. Each can be a powerful research technique for solving complex problems, and each has been adapted to a variety of needs through variants on the classical processes. Both processes require significant time commitments, and both are subject to bias. Both tend to discredit extreme positions and could alienate those stakeholders.

Q-Sort, on the other hand, is used primarily as an individual technique for developing theory related to human behavior and for identifying and describing

human phenomena. The Q-Sort is a time-consuming process, as are Delphi and NGT. In contrast to those methods, Q-Sort researchers develop an instrument first, through literature review. The instrument is designed to measure using forced-choice options. Data collection is usually accomplished one-on-one. Table 1 depicts similarities and differences between these three research methods.

Summary

The Delphi technique was designed to identify the best solutions to complex organizational and other social problems; and researchers in diverse fields have used it in its conventional form and with several variants. However, the process is fraught with opportunities for contamination through bias, either actual or perceived. Necessary as it is in a pluralistic society, both the possibility and the propriety of reaching consensus remains illusive.

—Ernest W. Brewer

See also Decision Theory

Further Reading

Adams, S. J. (2001, October). Projecting the next decade in safety management: A Delphi technique study. *American Society of Safety Engineers*, 32, 26–29.

Table 1 Comparisons and Contrasts: Q-Sort, Delphi, and Nominal Group Technique

<i>Name</i>	<i>Purpose</i>	<i>Data Collection</i>	<i>Primary Uses</i>	<i>Advantage</i>	<i>Disadvantage</i>
Delphi	Consensus building	Group; anonymous	Medicine; social sciences	Divergent thinking; does not require panel participants to meet	Possible manipulation
NGT	Decision making	Group; anonymous	Social sciences	Divergent thinking	Possible manipulation; requires participants to meet
Q-Sort	Theory building; description	Individual; forced choice	Psychology; social sciences	Quantified subjective data	Generalizability difficult

- Clayton, M. J. (1997). Delphi: A technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology*, 17(4), 373–386.
- Frazer, G. H., & Sechrist, S. R. (1994). A comparison of occupational stressors in selected allied health disciplines. *Health Care Supervisor*, 13(1), 53–65.
- Olmstead-Schafer, M., Story, M., & Haughton, B. (1996). Future training needs in public health nutrition: Results of a national Delphi survey. *Journal of the American Dietetic Association*, 96, 282–283.
- Schuler, R. (1995). *Managing human resources* (5th ed.). New York: West.
- Strauss, H. J., & Zeigler, L. H. (1975). The Delphi technique and its uses in social science research. *Journal of Creative Behavior*, 9, 253–259.
- Tavana, M., Kennedy, D. T., & Joglekar, P. (1996). A group decision support framework for consensus ranking of technical manager candidates. *Omega, International Journal of Management Science*, 24, 523–538.
- Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology? *Journal of Extension*, 35(5). Retrieved December 13, 2001, from <http://www.joe.org/joe/1997october/tt2.html>
- Michigan State University Extension. (1994, October). *Delphi technique*. Issue Identification Information – III00006. Retrieved November 28, 2001, from <http://www.msue.msu.edu/msue/imp/modiii/iii00006.htm>

Applying Ideas in Statistics and Measurement

The following abstract is adapted from Kelly, K. P. (2005). A survey of pediatric oncology nurses' perceptions of parent educational needs. *Journal of Pediatric Oncology Nursing*, 22(1), 58–66.

Educating parents of children with cancer is a primary nursing responsibility in pediatric oncology. Katherine Kelly used **Delphi techniques** with nurses attending a Children's Oncology Group Nursing Workshop to identify priority educational topics from pediatric oncology nurses' perspective. Nurses were asked to identify five priority educational topics and five topics on which they spend the most time teaching parents. Twenty-four educational categories were identified by 199 nurses, and responses were sorted by category and frequencies tabulated. Information about treatment was the most frequently cited priority. Bone marrow suppression was the second most important priority and was the topic

on which nurses spent the most time. In Round 2 of data collection, 132 consenting participants from Round 1 were asked to rate the importance of the categories from Round 1 (presented in random order) during four time periods (diagnosis, initial treatment, maintenance, and off therapy). Nurses reported different teaching priorities across the continuum of treatment. Interestingly, teaching about end-of-life issues and alternative therapy were ranked as low in importance across all time points.

DELTA METHOD

The Delta method allows one to find the approximate distribution of a function of a random variable. Often, we are interested in the variability or asymptotic distribution not of the random variable X directly, but rather of a function of that random variable, call it $f(X)$. It is usually not easy to calculate characteristics of $f(X)$ exactly; indeed, in many cases, it is impossible or nearly so, hence the appeal to approximations. The Delta method is one of the standard approaches.

The Delta method is based on Taylor series expansions from standard calculus, which we now briefly review. Suppose we have a function $g(y)$ for which derivatives exist up to order k . The Taylor series expansion of $g(y)$ about the point a is

$$g(y) = g(a) + \frac{g'(a)}{1!}(y-a) + \frac{g''(a)}{2!}(y-a)^2 + \dots + \frac{g^{(n-1)}(a)}{(n-1)!}(y-a)^{n-1} + \dots,$$

where we can continue taking derivatives of $g(y)$ as long as they exist. In practice, of course, it will always be necessary to stop the expansion after a finite number of terms, which leads to Taylor's formula with remainder. The remainder converges to zero rapidly enough that it is negligible compared to the rest of the series expansion and can be ignored.

For statistical purposes, we usually need for only the first derivative to exist, giving the approximation $g(y) = g(a) + g'(a)(y - a) + \text{remainder}$. Casting this in a more statistical light, suppose that y is a random variable with expected value θ and variance σ^2 . We expand $g(y)$ in a first-order Taylor series around θ , ignoring the remainder term, to obtain $g(y) \approx g(\theta) + g'(\theta)(y - \theta)$. If we take expectations on both sides of this last equality, we have

$$E[g(y)] \approx g(\theta) + g'(\theta)E(y - \theta).$$

Because the expected value of y is θ , the second summand drops out and we are left with $E[g(y)] \approx g(\theta)$.

Approximating the variance is equally easy, because

$$\begin{aligned} \text{Var}[g(y)] &= E[(g(y)) - g(\theta)]^2 \\ &\approx E[(g'(\theta)(y - \theta))^2] \\ &= [g'(\theta)]^2 \text{Var}(y) \\ &= \sigma^2 [g'(\theta)]^2. \end{aligned}$$

With these in hand, we can derive a Central Limit Theorem for the function $g(y)$. This extension of the basic Central Limit Theorem is the Delta method. Suppose $X_1, X_2, \dots, X_n$ is a sequence of random variables, and let T_n be a statistic based on the data such that $\sqrt{n} [T_n - \theta]$ converges in distribution to a normal with mean 0 and variance σ^2 . Then $\sqrt{n} [g(T_n) - g(\theta)]$ converges in distribution to a normal with mean 0 and variance $\sigma^2 [g'(\theta)]^2$, if $g'(\theta)$ exists and is not zero.

The result is best demonstrated via examples. In the first, let $X_1, X_2, \dots, X_n$ be independent, identically distributed $N(\theta, \sigma^2)$ and the parameter of interest is θ^2 . A reasonable estimator of θ is the sample average, $\bar{X}$, and a reasonable estimator of θ^2 is therefore the sample average squared, or $\bar{X}^2$. Here, $g(y) = y^2$, which yields $g'(y) = 2y$, so that $\sqrt{n} [\bar{X}^2 - \theta^2]$ converges in distribution to a normal with mean 0 and variance $4\theta^2\sigma^2$.

As a second example, let $X_1, X_2, \dots, X_n$ be independent, identically distributed Bernoulli (zero-one) trials, with probability p of success. Instead of p itself, we are interested in the odds of success, given by

$p/(1 - p)$. We will typically estimate p by $\hat{p} = \bar{X}$, the sample average, which in this case is the proportion of successes in the sample. Now, $g(y) = y/(1 - y)$ and $g'(y) = 1/(1 - y)^2$. Hence, our estimator $\hat{p}/(1 - \hat{p})$ is asymptotically normal, with mean $p/(1 - p)$ and variance $[g'(p)]^2 \text{Var}(\hat{p}) = p/n(1 - p)^3$.

The statement of the Delta method specifies that $g'(\theta)$ is not zero. If $g'(\theta)$ happens to equal zero, a modification of the result is necessary. This modification is simple: instead of taking a first-order Taylor expansion, we include the second-order term, that is $g(y) = g(a) + g'(a)(y - a) + g''(a)(y - a)^2/2 + \text{remainder}$. Because the term involving the first derivative vanishes, we now have, again expanding around θ , that $g(y) \approx g(\theta) + g'(\theta)(y - \theta)^2/2$. Moving $g(\theta)$ to the left-hand side of the equation yields $g(y) - g(\theta) \approx g''(\theta)(y - \theta)^2/2$. The square of a standard normal distribution is χ^2 with 1 degree of freedom, leading to the modification of the basic Delta method: Suppose $X_1, X_2, \dots, X_n$ are independent, identically distributed random variables, and T_n is a function of the data such that $\sqrt{n} [T_n - \theta]$ converges in distribution to a normal with mean 0 and variance σ^2 . Then $n[g(T_n) - g(\theta)]$ converges in distribution to $\sigma^2 \frac{g''(\theta)}{2} \chi_1^2$, provided that $g''(\theta) \neq 0$.

Returning to the first example, suppose that $\theta = 0$. Then $g(\theta) = \theta^2 = 0$, and the standard Delta method cannot be applied. We can use the modification to study the distribution of $\bar{X}^2$ in this instance, because $g'(\theta) = 2\theta = 0$ when $\theta = 0$, but $g''(\theta) = 2$ no matter what the value of θ . We can then conclude that when $\theta = 0$, $n[\bar{X}^2 - \theta^2]$ converges in distribution to $\sigma^2 \frac{2}{2} \chi_1^2 = \sigma^2 \chi_1^2$.

A second modification of the basic method involves a function of multiple parameters that is estimated by a function of more than one random variable. The underlying theory stems from multivariate calculus and a multivariate version of the Taylor series expansion, but the essence of the technique is the same, namely, expand around the parameters to terms up to first order (i.e., partial first derivatives of the function with respect to each of the parameters). From this expression, we can derive the approximate mean and variance of the function for use in an asymptotic normal distribution. It is possible to extend the result even further by considering more than one function of

the parameters simultaneously, leading to a multivariate normal distribution in the Delta method approximation.

The modification to allow for functions of random variables is particularly useful for ratio estimators, which arise frequently in practice but can be difficult to handle theoretically. Again, a simple example is instructive. Suppose $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$ are samples of independent, identically distributed random variables, where $E(X) = \theta_x$, $E(Y) = \theta_y$, $\text{Var}(X) = \sigma_x^2$, $\text{Var}(Y) = \sigma_y^2$, and $\text{Cov}(X, Y) = \sigma_{xy}$. θ_x and θ_y are both nonzero. The function of interest is $g(\theta_x, \theta_y) = \theta_x/\theta_y$. To proceed, first take the partial first derivatives of $g(\theta_x, \theta_y)$ with respect to θ_x and θ_y :

$$\frac{\partial}{\partial \theta_x} \frac{\theta_x}{\theta_y} = \frac{1}{\theta_y}$$

and

$$\frac{\partial}{\partial \theta_y} \frac{\theta_x}{\theta_y} = \frac{-\theta_x}{\theta_y^2}$$

A natural estimator for $g(\theta_x, \theta_y)$ is $\bar{X}/\bar{Y}$, the ratio of the sample averages. Then, by the same reasoning as in the univariate case,

$$E\left(\frac{\bar{X}}{\bar{Y}}\right) \approx \frac{\theta_x}{\theta_y}$$

and

$$\begin{aligned} \text{Var}\left(\frac{\bar{X}}{\bar{Y}}\right) &\approx \frac{1}{\theta_y^2} \text{Var}(\bar{X}) + \frac{\theta_x^2}{\theta_y^4} \text{Var}(\bar{Y}) \\ &\quad - 2 \frac{\theta_x}{\theta_y^3} \text{Cov}(\bar{X}, \bar{Y}). \end{aligned}$$

These can then be used to obtain the approximate normal distribution of $\bar{X}/\bar{Y}$.

When competing estimators of the same parameter or function of the parameter are available, the Delta method provides a convenient way of comparing them, because one of its by-products is an estimate of variability.

—Nicole Lazar

See also Normal Curve; Random Sampling

Further Reading

- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.
 Lehmann, E. L. (1991). *Theory of point estimation*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

DEMING, WILLIAM EDWARDS (1900–1993)

At the time of his death in 1993, Ed Deming was regarded as a world leader in quality management; he had been voted by the business staff of the *Los Angeles Times* as being one of the 50 most influential businesspeople of the century. His obituary in the *American Statistician* was headed “The Statistician Who Changed the World.” However, Deming simply described himself as a “consultant in statistical studies.”

Deming was born to a poor family in Sioux City, Iowa, on October 14, 1900. His mother was a music teacher, and Deming had a lifelong interest in music (he played the flute). He studied electrical engineering at the University of Wyoming, graduating in 1921. He followed this with an MS in mathematics and physics at the University of Colorado in 1925. As a summer job, he worked for the Western Electric Company in Chicago, where he first encountered Shewart’s work on quality control. He obtained his doctorate in physics from Yale University in 1928.

Deming began working first for the U.S. Department of Agriculture and then for the U.S. Bureau of the Census, being responsible for the sampling methods used for the first time in the 1940 U.S. Census. Between 1930 and 1946, Deming was a special lecturer on mathematics and statistics in the graduate school of the National Bureau of Standards, giving lectures from 8 a.m. to 9 a.m. These led, in 1947, to the establishment of the Statistical Engineering Laboratory within the Bureau of Standards.

In 1946, Deming began his practice as a statistical consultant. In 1947, he spent three months in Japan helping with the Japanese census. On his return to Japan in 1950, he gave an extended course in quality control. The course was so successful and influential that he was invited back on many occasions and received by Emperor Hirohito. In 1960, Deming was

awarded the Second Order of the Sacred Treasure. At that time, he was much better known in Japan than in his home country.

Deming's 14 key principles for transforming business effectiveness are summarized thus:

1. Create constancy of purpose.
2. Take the lead in adopting the new philosophy.
3. Cease dependence on mass inspection to achieve quality.
4. End the practice of awarding business on the basis of cheapness.
5. Improve constantly.
6. Institute training on the job.
7. Institute leadership.
8. Drive out fear and build trust.
9. Break down barriers between departments.
10. Eliminate slogans, exhortations, and targets.
11. Eliminate numerical goals, and management by objective. Substitute leadership.
12. Remove barriers to pride in workmanship.
13. Institute a program of education and self-improvement.
14. Put everybody to work to accomplish the transformation.

Deming was President of the Institute of Mathematical Statistics in 1945. In 1983, he was awarded the Wilks medal, the highest honor of the American Statistical Association. In 1987, he was awarded the National Medal for Technology. Deming died in Washington, DC, on December 20, 1993.

—Graham Upton

Further Reading

Walton, M. (1986). *The Deming management method*. New York: Perigee.

W. Edwards Deming Institute: <http://www.deming.org/>

in scientific research. The dependent variable, sometimes referred to as the *dependent measure*, *criterion variable*, or *Y variable*, is the experimental variable that is measured to determine the effects of an independent variable (e.g., experimental treatment) on selected subjects during a research experiment. Using the subjects' performance on the dependent variable, the researcher attempts to determine a relationship between the independent variable that is manipulated and outcomes on one or more dependent measures. For example, in a study of how the implementation of a newly designed reading program is related to improved scores on reading comprehension of fourth-grade students, the students' performance on the reading comprehension test after they were taught using the new reading program is the dependent variable, whereas the independent variable, the manipulated variable, is the reading program. *Dependent variable* is a generic term that can encompass many different types of measurements. Examples of dependent variables often seen in the research literature are posttest, transfer test, generalization test, probes, unit tests, and so on. The researcher typically reports the results of the study in table or graphic form. Regardless of the type of research design, the dependent variable is a necessary feature of the research.

Characteristics

A dependent variable, one in which both the researcher and consumer can have confidence, should have certain characteristics. Thus, in any well-designed research study, the dependent variable should be (a) clearly defined; (b) closely linked to the independent variable; (c) reliable and valid; (d) sensitive to treatment effects; and (e) administered by the researcher under prescribed, carefully monitored procedures.

Research Examples

Although the number and types of dependent variables are innumerable, several examples are presented across five specific, frequently used research methodologies.

DEPENDENT VARIABLE

The term *dependent variable* is derived from mathematics and is basic to understanding results

Survey

Research question: What is the average income of physicians in the United States?

Dependent variable: Income of all physicians in the United States.

Independent variable: None.

Correlation

Research question: Are years of education related to income?

Dependent variable: Income of workers across all education levels.

Independent variable: Years of education of the research sample.

Experimental

Research question: Does a certain assertiveness training program help salespersons earn more money?

Dependent variable: Income of selected salespersons.

Independent variable: Assertiveness training program.

Naturalistic-Observational

Research question: What is the frequency of the use of punishment procedures by ninth-grade resource room, special education teachers during class?

Dependent variable: The frequency of punishment used by ninth-grade resource room teachers during class.

Independent variable: None.

Single-Subject Reversal Design

Research question: Does teacher reinforcement during instruction increase the attending behavior of a child with mild mental retardation?

Dependent variable: Percentage of instructional time the student is attending during reinforcement and teaching condition.

Independent variable: Use of teacher reinforcement during instruction.

—Craig Darch and Ronald C. Eaves

See also Independent Variable

Further Reading

Keppel, G., Saufley, W., & Tokunaga, H. (1992). *Introduction to design and analysis: A student's handbook*. New York: W. H. Freeman.

Research methods in the social and natural sciences: http://www.mcli.dist.maricopa.edu/proj/res_meth/

DESCRIPTIVE RESEARCH

Descriptive research provides a detailed account of a social setting, a group of people, a community, a situation, or some other phenomenon. This kind of research strives to paint a complete and accurate picture of the world by focusing on the factual details that best describe a current or past event. Researchers engaged in descriptive studies set out to identify who participates in an event, where and when it occurs, and what happens, without exploring the causal relationships involved in that event. For example, a descriptive study may examine the types of services offered by a government agency, the living conditions of a homeless population in a large urban center, the experiences of teachers in elementary school classrooms, or the daily needs of individuals living with breast cancer. One common example of a descriptive study is a census, which sets out to document demographic (e.g., age, gender) and other details (e.g., housing costs) about individuals living in a particular community. Census data are often collected over many years, allowing researchers to examine changes in demographic and social patterns within a particular nation, city, neighborhood, or other identified social grouping.

In compiling descriptive facts about various phenomena, descriptive research is allied most closely with quantitative approaches (including the use of descriptive statistics), although descriptive approaches may also be used in qualitative research to provide valuable background information for analyses of individuals' attitudes, opinions, and personal experiences of particular phenomena. Descriptive research is the most commonly used approach in the human (behavioral) sciences because it allows researchers to

examine conditions that occur naturally in the home, hospitals, classrooms, offices, libraries, sports fields, and other locales where human activities can be systematically explored, documented, and analyzed.

Descriptive Research Methods

In quantitative research, descriptive studies are concerned with the functional relationships between variables, hypothesis testing, and the development of generalizations across populations. The findings of descriptive studies are valuable in that they provide information that enables researchers and practitioners to define specific variables clearly, to determine their current situations, and to see how these variables may relate to other variables. In qualitative approaches, descriptive research is often referred to as a form of naturalistic inquiry; this type of research allows the researcher to observe, document, and detail specific activities within a defined social setting in order to point to transferable findings. In both quantitative and qualitative approaches, descriptive research is marked by its exploration of existing events and conditions that would have happened even if the researcher was not there to observe and document the details. A number of different research methods are commonly used in quantitative and qualitative descriptive studies; the sections that follow will briefly examine the goals of some of these approaches.

Questionnaires and Structured Interviews

Methods designed to survey individuals about their experiences, habits, likes and dislikes, or even the number of televisions in their homes are commonly used to gather data from a large sample of a given population at a particular point in time. These methods are designed to generalize to the larger population in order to document the current or past activities and experiences that surround a particular phenomenon. For example, a questionnaire may be designed to identify young people's familiarity with different media outlets, to explore parents' knowledge about treatments for the common cold, or to document the demographic characteristics of new immigrants in

rural communities. Large-scale questionnaires and structured interviews typically use some form of probability sampling to select a representative sample of a particular population. These methods take many different forms and can be used across topic areas, including telephone polls (e.g., to solicit voting patterns), mail-in or Web-based questionnaires (e.g., personal shopping habits), and in-person surveys (e.g., in-store product assessments). Researchers must take care to ensure high response rates that will represent the population, as participation rates as low as 15% can be common, especially in e-mail or Web-based surveys.

One of the most common examples of this type of research is an opinion poll, which is typically designed to document demographic details about individuals (e.g., their highest level of education) as well as their opinions on such topics as children being required to wear uniforms in schools, Internet use in the home, mass media as a source of health information, or other issues of social relevance. Question response types may include yes/no, multiple choice, Likert scale, open-ended (short answer) questions, or other appropriate designs. The results of such polls are typically analyzed with fairly simple techniques designed to organize and summarize the findings, such as the calculation of the mean number of women versus men in favor of capital punishment.

Observation

Observing human behavior in natural settings (such as watching shoppers as they stand in line at the grocery store, or patients as they sit in an emergency waiting room) can elicit insightful data that could not be captured using other data collection methods. The data gathered using observational approaches consist of detailed descriptions of people's activities and behaviors, as well as physical details about the social settings that surround and inform those activities. Observational techniques may be covert or overt, and may even result from a researcher's involvement in the particular social scene being investigated (e.g., librarian researchers who work at the public library's reference desk); this latter technique is

known as participant observation. Researchers employing observational methods not only document details about the individuals within the setting under study (say, in an emergency room), but also examine the physical (e.g., location of triage facilities) and organizational (e.g., management hierarchies) structures within that setting. Data collection may be restricted to a single site (e.g., one classroom) or may involve multiple sites (e.g., all classrooms within all schools in a district), but typically extends over a long period of time in order to gather valid and complete data.

One example of an observational technique used for gathering information about individuals in a particular social context is the “seating sweeps” method that was developed for use in a public library context. This method involved the use of checklists to document basic demographic information about library patrons (e.g., gender); the activities in which they were engaged (e.g., computer use, reading); where they engaged in those activities (e.g., private study carrels, computer lab); and the materials that these patrons carried with them (e.g., briefcases, writing materials). A number of general patterns emerged about human behavior in the library using this observational technique, including the number of men who used the library at various times of the day and week, and the prevalence of personal entertainment devices used by library patrons.

In-Depth Interviews

In-depth interviews allow researchers to examine issues at length from the interview respondent’s personal perspective, and they are commonly used in qualitative research approaches. The data gathered during interviews typically consist of verbatim responses to the interviewer’s questions, which are designed to elicit descriptions of personal behaviors, and the opinions, feelings, and attitudes that inform those behaviors. Interviews typically last from 60 to 90 minutes, although the length varies depending on the scope of the project and the availability of participants. Common themes and patterns that emerge from the data derived from these interviews can guide researchers in the assessment of existing programs

and services and in the exploration of various social issues. Transferable findings generally occur at the point of saturation of themes in the data, which typically arise with a minimum of 15 to 18 participants. Increasing the number of interviewees is one way to enhance rigor in data collection and to speak more authoritatively about the findings under study. However, it is also worth noting here that anomalies in the data (such as the experiences of a single individual who provides details about an experience that is unlike that of other interviewees) can also be extremely valuable to qualitative researchers. These singular experiences can highlight individuals’ particular needs, especially in settings where policies and practices have been designed for majority populations, and often point to areas that require additional research. In-depth interviews can also be combined with other methods (e.g., structured computer tasks used to assess Web site usability, quantitative questionnaires designed to elicit factual data) to provide a more complete picture of the phenomena under study.

Focus Groups

Focus groups also fall into the interview category and may be either highly structured (i.e., quantitative) in nature, or designed to be more of a personal dialogue between participants (i.e., qualitative). In either case, the defining feature of these interviews is that they occur with groups of individuals (typically five to eight people, with one or more groups in total) whose comments are focused on a particular issue of interest to the researcher. Participants are typically fairly homogeneous group members (e.g., new immigrants living in a particular city, undergraduate students using campus recreational facilities) who are asked to reflect on a series of questions or to react to new products or policies. These interviews can be more challenging to conduct than individual interviews because of the need to manage group dynamics (e.g., ensuring that all group members are able to speak their minds without feeling silenced by other group members). These interviews are best run by a trained facilitator, often require a more formal setting (such as a boardroom), and may take more time to coordinate than other survey methods.

Personal Journals and Diaries

Asking individuals to document their daily activities (such as when or how often they have used an organization's Web site) can be an effective way to document human behavior. One benefit of this approach is that data are collected as they happen, so that researchers need not rely on the accuracy of individuals' memories of events (as in interviews, questionnaires, or other methods where individuals are asked to discuss their behaviors). For example, this method can be used by physicians to track patients' meals and other activities related to personal health, or by education researchers to track students' study habits. Personal journals and diaries allow individuals to document quantitative elements of their activities (such as how often they go to the grocery store and how much money they spend per trip), as well as their thoughts, feelings, and experiences of shopping in particular stores or for particular items. Participants typically need some instruction in the researcher's expectations (e.g., how much detail to provide, how often to write an entry, what topics to include), but can often provide much more detail than is possible to gather using an interview or other research method. Individuals may keep diaries for a period of a week or more, and may write on a variety of topics, which can then be examined further with other, follow-up methods (e.g., personal interviews).

Whether used on their own or in conjunction with one another, all of these methods are useful tools for gathering data on various elements of human behavior. Descriptive research provides valuable insight into the social scenes that surround and inform our lives. The knowledge that we gain about social settings, people, specific experiences and activities, and other elements of social behavior are useful to practitioners (such as hospital and school administrators, or government officials), but also inform other research approaches. Descriptive research can act, for example, as a first step in a more detailed and complex study of social behavior, providing valuable background details about individuals or information on variables that require more advanced study. However, descriptive studies also stand in their own right as a means to examine,

document, and reflect on the world and illuminate the social phenomena that inform individuals' personal and work-related lives.

—Lisa M. Given

See also Inferential Statistics

Further Reading

- Given, L. M., & Leckie, G. J. (2003). "Sweeping" the library: Mapping the social activity space of the public library. *Library & Information Science Research*, 25(4), 365–385.
- Ruane, J. M. (2005). *Essentials of research methods: A guide to social science research*. Malden, MA: Blackwell.
- Sarafino, E. P. (2005). *Research methods: Using processes and procedures of science to understand behavior*. Upper Saddle River, NJ: Pearson-Prentice Hall.

DEVIATION SCORE

The deviation score is the difference between a score in a distribution and the mean score of that distribution. The formula for calculating the deviation score is as follows:

$$X - \bar{X}$$

where

$\bar{X}$ (called "X bar") is the mean value of the group of scores, or the mean; and

the X is each individual score in the group of scores.

Deviation scores are computed most often for the entire distribution. For example, for the following data set (see Table 1), there are columns representing scores on the variables X and Y for 10 observations. The deviation scores for X and Y have also been calculated. Notice that the means of the deviation score distributions are zero.

Thus, the deviation scores are simply a linear transformation of a variable. This can be demonstrated by calculating the Pearson correlations between X and Y and then between the deviation- X and deviation- Y scores. In both instances, the correlations are 0.866.

Table 1 Raw and Deviation Scores on Two Variables, X and Y

Observation	X	Y	$X - 4.8$	$Y - 4.2$
1	2	1	-2.8	-3.2
2	3	4	-1.8	-0.2
3	4	3	-0.8	-1.2
4	7	5	2.2	0.8
5	8	6	3.2	1.8
6	9	8	4.2	3.8
7	2	3	-2.8	-1.2
8	3	3	-1.8	-1.2
9	4	2	-0.8	-2.2
10	6	7	1.2	2.8
	$\bar{X} = 4.8$	$\bar{Y} = 4.2$	$\bar{X} = 0.0$	$\bar{Y} = 0.0$

The next question one might want to ask is, Why would one want to calculate such scores? The most frequent use of deviation scores is in conducting simultaneous solution regression analyses when there is an interest in the effects of interaction terms.

For example, assume one wants to predict a criterion (Z) with two main effects, X and Y , as well as their interaction. The interaction term is generated by multiplying X and Y , but this interaction term exhibits multicollinearity with each of the main effects, X and Y . However, if the interaction term is created from the deviation scores of X and Y , the multicollinearity no longer is a problem.

Table 2 Raw Scores and Interaction Terms for Nondeviation and Deviation Scores

Observation	X	Y	$(X)(Y)$	$X - 4.8$	$Y - 4.2$	$(X - 4.8)(Y - 4.2)$
1	2	1	2	-2.8	-3.2	8.96
2	3	4	12	-1.8	-0.2	0.36
3	4	3	12	-0.8	-1.2	0.96
4	7	5	35	2.2	0.8	1.76
5	8	6	48	3.2	1.8	5.76
6	9	8	72	4.2	3.8	15.96
7	2	3	6	-2.8	-1.2	3.36
8	3	3	9	-1.8	-1.2	2.16
9	4	2	8	-0.8	-2.2	1.76
10	6	7	42	1.2	2.8	3.36

To demonstrate this, the data set shown earlier is used (see Table 2). The interaction terms have been generated for each score. The correlations between the nondeviation interaction, $(X)(Y)$, and the main effects are .955 with X and .945 with Y . The correlations between the deviation interaction $(X - 4.8)(Y - 4.2)$ and the main effects are .479 with X and .428 with Y . This feature of deviation scores is of immense utility when conducting simultaneous linear regression-based analyses (such as multiple regression, discriminant function analysis, logistic regression, and structural equation modeling).

—Theresa Kline

See also Standard Deviation; Standard Scores; Variance

Further Reading

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Kline, T. J. B., & Dunn, B. (2000). Analysis of interaction terms in structural equation models: A non-technical demonstration using the deviation score approach. *Canadian Journal of Behavioural Science*, 32, 127–132.

DIAGNOSTIC VALIDITY

Diagnostic validity applies to any test, measurement, or decision-making strategy that categorizes people.

Also referred to as *categorical validity* or, more pragmatically, as the 2×2 table, diagnostic validity examines the relationship between how a test categorizes a subject and in which category the subject actually is. Relevant categories might include, among others, HIV-positive individuals, top employment prospects, violent recidivists, child molesters, fit parents, suitable graduate students, or incompetent defendants. Validity information answers questions regarding the probability that a classification is correct, the utility of the

test or strategy for different purposes, and how to interpret the classification. This information also solves Bayes' theorem: We often know the percentage of paranoids, say, who score positively on a test of paranoia (by administering the test to a large group of paranoids); Bayes' theorem computes the reasonableness of inferring paranoia from a positive test score. The answer requires knowledge about the incidence of paranoia and about how nonparanoids do on the test.

In this entry, *test* is used specially to mean any score, sign, symptom, or series of these used to categorize people. The *Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV)* is a manual of tests. Each diagnosis is accompanied by a test to determine if a particular subject has the diagnosis in question. For example, the test for paranoid personality is (a) the presence of a personality disorder, plus (b) the presence of at least four of seven behaviors, plus (c) the exclusion of some other diagnoses. The fact that this is a test for paranoid personality is disguised by the failure of the publishers to include the 2×2 table that would answer questions about diagnostic validity. The test is made to look like the *definition* of the disorder instead of a method of detecting who has the disorder and who does not. Before the test was codified, there must have been some other way to determine who had paranoid personality disorder and who did not, and that method was extremely unlikely to be a perfect fit with the current criteria, even if that other way was only in the imagination of the test writers.

As an example, let's assume that the *DSM-IV* test for paranoid personality was a very good test. (We can only assume because the actual data have not been published.) Assume that some expert clinicians carefully identified 100 people as having this disorder. Further assume that the test published in *DSM-IV* gave a positive result for all 100 individuals, and that of 100 randomly selected psychiatric admissions without paranoid personality disorder, only 5 tested positive with the *DSM-IV* criteria. That would certainly be impressive, 100 out of 100 correctly identified with the disorder, and 95 out of 100 correctly identified without the disorder.

Now for some terminology. In this case, the expert clinicians' original diagnoses constitute the *gold standard*, which is the method by which subjects were placed in their actual categories. The gold standard is crucial for interpreting test results, because even the best test predicts only the categories assigned by the gold standard. Thus, for example, a test of violent recidivism usually has a gold standard of rearrest as the indicator of recidivism, so the test can never be better at identifying recidivists than rearrest is, and it is obvious that there are some people who recidivate but are not caught. Furthermore, there are some people who *are* arrested, but incorrectly, and not because they recidivated. Understanding the gold standard is crucial to understanding what a test that categorizes people is able to achieve.

Sensitivity is the accuracy of the test among people who have the condition (who are actually in the category). In our paranoid personality example, the sensitivity is 100/100, or 1.00. *Specificity* is the accuracy of the test among people who do not have the condition (who are not actually in the category). In this case, the specificity is 95/100. *True positives* (TP) are those subjects who are in the category and who are identified as such by the test. *True negatives* (TN) are correctly identified by the test as not being in the category. *False positives* (FP) are not actually in the category but the test says they are; *false negatives* (FN) are actually in the category, but the test says they are not. Sensitivity = $TP/(TP + FN)$. Specificity = $TN/(TN + FP)$.

The *cutoff score* is the decision point at which test results are considered positive or negative for being in the category. In the *DSM-IV*, the cutoff score for paranoid personality is a combination of meeting the first and third criteria, plus four of the behaviors listed in the second criterion. On a typing test to determine good employees, the cutoff score might be 60 words per minute; for graduate students administered the Graduate Record Exam, it might be 1200. Each cutoff score produces a different 2×2 table for the analysis of diagnostic validity. A statistic called the area under the curve (AUC) of the receiving operator characteristic curve (ROC) can be computed that evaluates the test as a whole, independent of the cutoff score. The AUC expresses the probability that a person in the

category will score higher on the test than a person not in the category. For categorization, though, a cutoff score must be selected and analyzed in the 2×2 table (see Table 1).

Positive predictive value (PPV) is the accuracy of the test among people who test positive. It tells us how seriously to take a positive result. $PPV = TP/(TP + FP)$. *Negative predictive value* (NPV) is the accuracy of the test among people who test negative. It tells us how seriously to take a negative result. $NPV = TN/(TN + FN)$.

The *hit rate* of a test is the overall percentage of accurate classifications, or $(TP + TN)/(TP + FP + FN + TN)$. The value of a test is, generally, in the improvement it produces over and above the hit rate that would be obtained from assigning everyone to the more populous category. An exception would be a test used for screening rather than for categorization. If a cheap, harmless medication were available for a serious disease, a desirable test would have perfect sensitivity, so that everyone with the disease got the treatment. But if the specificity were mediocre, that would be fine, because there would be little downside to overadministering the medication.

Before judging a test's validity, we need to know or estimate the *base rate* of the condition or category. The base rate is the incidence of the category among relevant subjects. If the *DSM-IV* test for paranoid personality were being used in employment screening, then the relevant base rate would be the incidence of

this disorder in the general population. If used to diagnose psychiatric patients, then the base rate would be the incidence among hospital admissions. For the sake of this discussion, assume that the incidence of paranoid personality is 0.5% in the general population, and 2% in psychiatric admissions. Because the specificity data (95/100) were obtained from 100 randomly selected psychiatric admissions, we have the data needed to analyze this as a test only in that context. Therefore, in filling in the 2×2 table, we need to adjust the column of persons without the condition so that they are represented according to their base rate (98%), while maintaining the test's specificity of 95%.

Notice that the numbers in Table 2 preserve the sensitivity of the test (100% of persons with the disorder are correctly classified); its specificity (95% of persons without the disorder are correctly classified); and the base rate (the table reflects a population in which 2% of persons, or 100 out of 5,000, have the disorder). The PPV is 100/345, meaning that there is only a 29% chance (given our assumptions) that a person diagnosed with paranoid personality disorder using the *DSM-IV* actually has the disorder. Such is the fate of trying to classify people into categories with low base rates. Because the base rate is, under our assumptions, only 2%, we could correctly classify 4,900 out of 5,000 admissions by simply claiming that none of them has the disorder. Even this excellent test, with its 100% sensitivity and 95% specificity,

Table 1 The Basic 2×2 Table

	G O S L T D D	
	Actually in the category	Actually not in the category
Test says in the category	TP	FP
CUTOFF SCORE		
Test says not in the category	FN	TN

Table 2 Hypothetical Example: A Test of Paranoid Personality Disorder

	G O S L T D D	
	Actually in the category	Actually not in the category
Test says in the category	TP = 100	FP = 245
CUTOFF SCORE		
Test says not in the category	FN = 0	TN = 4,655

Table 3 Hypothetical Example: A Racial Test of Violence Risk

	G O S L T D D	
	Actually in the category (violent intent)	Actually not in the category (no violent intent)
Test (black) says in the category	TP = 26	FP = 10,924
CUTOFF SCORE		
Test (not black) says not in the category	FN = 26	TN = 10,939,024

correctly classifies only 4,755 out of 5,000 admissions. In terms of Bayes' theorem, we have calculated the probability of A given B (the probability of having the disorder given a positive test score) from the probability of B given A (the probability of a positive test score given the disorder).

To demonstrate the breadth of the applicability of the 2×2 table, consider the case of *racial profiling*. Racial profiling is the statistical justification of police suspicion derived from an increased likelihood of criminal activity based on a suspect's race. In this hypothetical example, imagine a wealthy community whose police routinely stop black motorists. The police justify this conduct by noting that in this all-white community, 50% of all non-domestic violent crimes are committed by black people, while only 0.1% of cars observed in the town have black drivers. The category is *motorist intent on violent crime*, and the test is whether or not the motorist is perceived as black by the police. The probability of B (testing positive for looking black) given A (being a violent criminal) is 50%. What is required to justify stopping black motorists, though, is the probability that a black driver is a violent criminal, not the probability that a violent criminal is black. In other words, how good a test of criminality is being black under these circumstances?

To fill in the 2×2 table, we need to know or estimate the base rate of non-domestic violence perpetration

among car drivers. Suppose this community's streets convey 30,000 motorists a day, and there is one violent crime a week. In the course of the year, the community sees 52 crimes and 10,950,000 car trips. (See Table 3.)

Even though half of all non-domestic violent crimes are committed by black people and black people account for only 0.1% of car trips, the probability of a black driver being a violent criminal in this scenario is only 26/10950. Two chances in a thousand does not justify a police stop. The low base rate of violent crime in this community makes the test of race a useless one, regardless of its consequences for social justice.

Special Problems With the Gold Standard

Data regarding test validity depend on the original sample that was separated into groups by the gold standard. Any sample may have idiosyncratic or unexpected features. For example, the *DSM-IV* test for paranoid personality includes an item about bearing grudges for insults, but it is conceivable that such grudges are a feature of paranoia only in some subcultures. For this reason, every test that categorizes people should be cross-validated on a separate sample. Cross-validation does not guarantee elimination of idiosyncrasies in the sample (or in the employment of the gold standard), because two samples may have the same idiosyncrasy, but lack of cross-validation makes the presence of an idiosyncrasy too likely for an uncrossed test to be trusted.

Care must be taken not to conflate the gold standard with test items, so as to avoid the creation of a pseudo-test. In the example above that deals with racial profiling, it is unclear what the gold standard was for determining which crimes were violent and which were not. Conceivably, the perceived race of the defendant might have influenced this determination. Then, to include race of the defendant as a test item for predicting violence conflates the test and the gold standard, and is bound to make the test look better than it is. This is a form of the logical error, begging the question, or assuming the conclusion.

Certain gold standards are so subjective that tests validated against them cannot be separated from the people on whose subjectivity they depend. In certain contexts, this is not a problem, because the desired use of the test is to please the original judges. For example, employers may choose which are the good current employees and which are the undesirable ones, as long as it is understood that the resulting test is designed to select employees with whom the employer will be pleased, and not employees who meet some other criterion. Tests of good parenting, competency to stand trial (CST), and mental retardation (MR), on the contrary, cannot escape the arbitrariness of the gold standard used to categorize the original sample. Good parenting is obviously subjective. CST is a category that does not occur in nature, but only in the minds of judges. For a test to be useful, the test must be cheaper or more convenient than the gold standard it tries to approximate. Any test of CST is an attempt to improve on a classification that is simple and, by definition, nearly perfect (judges' classifications of CST are rarely overturned by appellate courts). MR also does not occur in nature, but instead represents the arbitrary and politically determined percentage of people whom the society thinks is too limited to be held accountable for self-care. There are too few real differences between subjects who score 69 on an IQ test (MR) and those who score 75 (not MR) to employ an objective gold standard for the validation of, say, a test of adaptive functioning that purports to distinguish people with and without MR.

—Michael Karson

See also Validity Theory

Further Reading

Mart, E. G. (1999). Problems with the diagnosis of factitious disorder by proxy in forensic settings. *American Journal of Forensic Psychology*, 17(1), 69–82.

Wood, J. M. (1996). Weighing evidence in sexual abuse evaluations: An introduction to Bayes' theorem. *Child Maltreatment*, 1(1), 25–36.

Area under the curve and receiving operator characteristic curve description: <http://www.anaesthetist.com/mnm/stats/roc/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Miller, J. D., Bagby, R. M., Pilkonis, P. A., Reynolds, S. K., & Lynam, D. R. (2005). A simplified technique for scoring *DSM-IV* personality disorders with the five-factor model. *Assessment*, 12(4), 404–415.

There are many different types of validity, with **diagnostic validity** being the one that examines how “true” or valid a set of diagnostic criteria is for a certain condition. The current study compares the use of two alternative methodologies for using the Five-Factor Model (FFM) to assess personality disorders (PDs). Across two clinical samples, a technique using the simple sum of selected FFM facets is compared with a previously used prototype matching technique. The results demonstrate that the more easily calculated counts perform as well as the similarity scores that are generated by the prototype matching technique. Optimal diagnostic thresholds for the FFM PD counts are computed for identifying patients who meet diagnostic criteria (used to help establish diagnostic validity) for a specific PD. These threshold scores demonstrate good sensitivity in receiver operating characteristics analyses, suggesting their usefulness for screening purposes. Given the ease of this scoring procedure, the FFM count technique has obvious clinical utility.

DIFFERENCE SCORE

The difference score indicates the amount of change between two testings. It is computed by subtracting the score on the first testing from the score on the second

$$d = Y - X,$$

where

d is the difference score (sometimes called *change score* or *gain score*),

X is the first test score (sometimes called the *baseline* or *pretest score*), and

Y is the second test score (sometimes called the *posttest score*).

In SPSS, difference scores are created by computing a new variable. This is done using the Compute function found under the Transform window. The syntax for computing a new variable called “change” to indicate the change from anxiety1 to anxiety2 would be as follows:

```
COMPUTE change = anxiety2 - anxiety1.
EXECUTE .
```

See the example in Table 1.

Table 1 Example of Difference Scores

ANXIETY1	ANXIETY2	CHANGE
23	20	−3.00
45	40	−5.00
26	23	−3.00
34	35	1.00
52	44	−8.00

In Table 1, four of the five participants showed a decrease in anxiety as indicated by the negative difference scores.

Difference scores can be treated like any other variable. The mean of difference scores equals the difference between the means from the two testings. In the above example, the mean of anxiety1 is 36.0, the mean of anxiety2 is 32.4, and the mean of the difference scores is −3.6. This shows that the average change from the first to the second testing was a *decrease* in anxiety of 3.6.

Advantages

Difference scores generally have much less variation than the scores from which they were created. This is because the subtraction operation removes any variation due to individual characteristics that is constant between the two testings. Thus, analyses using difference scores offer more statistical power than analyses conducted on posttest scores.

Difference scores allow a simpler design to be used. A one-way ANOVA comparing the means of difference scores yields a main effect that is identical

in both value and meaning to the interaction term in a two-way ANOVA that used the pretest and posttest scores as a second, repeated measures variable. Post hoc comparisons of the mean changes are easier to conduct and interpret in the one-way design.

Disadvantages

Difference scores contain measurement error from both the pretest and posttest scores, and are also negatively correlated with baseline because of measurement error. However, neither of these factors prohibits their use as valid measures of change.

On the other hand, when data are skewed—for example, by a floor or ceiling effect—difference scores may not reflect the true amount of change.

Appropriateness for Comparing Changes in Means

In a randomized experiment, where the goal is to compare the mean changes of groups that receive different treatments, analysis of covariance (ANCOVA), with pretest as the covariate and posttest as the dependent variable, should be used instead of difference scores. ANCOVA provides a better adjustment for minor differences in the pretest means because these differences are entirely due to chance and will regress on the second testing.

However, with naturally occurring groups, where the goal is to compare the changes of different groups to the same treatment, difference scores should be used instead of ANCOVA because the pretest differences between groups are not entirely due to chance and will not regress. ANCOVA would yield incorrect and directionally biased conclusions; for example, when scores are increasing from pretest to posttest, greater increases would generally appear for the group with the higher baseline.

Appropriateness for Examining Predictors of Change

Difference scores should be used instead of residual scores to study predictors (correlates) of change, because correlations between predictors and residual

scores are confounded by correlations between predictors and baseline. This is analogous to avoiding ANCOVA with naturally occurring groups. There is also an analogous directional bias; for example, residual scores are biased toward finding positive correlations with change for predictors that have positive correlations with baseline.

—John Jamieson

See also Dependent Variable

Further Reading

Salkind, N. J. (2007). *Statistics for people who (think they) hate statistics: The Excel edition*. Thousand Oaks, CA: Sage.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, 4(3), 265–287.

Difference scores are used in all kinds of studies, and even though their use is widespread, they suffer from numerous methodological problems. Jeffrey Edwards discusses how these problems can be avoided with polynomial regression analysis, a method that has become increasingly prevalent during the past decade. However, a number of potentially damaging myths have begun to spread regarding the drawbacks of difference scores and the advantages of polynomial regression, and if unchecked, difference scores and the problems they create are likely to persist. This article reviews 10 myths about difference scores and attempts to dispel these myths.

DIFFERENTIAL APTITUDE TEST

The Differential Aptitude Test (DAT), first published in 1947 by The Psychological Corporation, is a battery of tests whose goal is to assess multiple separate aptitudes of students and adults. The latest (fifth) version of the DAT, published in 1990, assesses verbal and numerical reasoning, mechanical reasoning, perceptual ability, spatial relations, abstract reasoning,

spelling, and language usage. Separate scoring norms are available for individual tests in the battery. The DAT is available in two levels: Level 1 of the DAT was designed for students in Grades 7 to 9 and adults who have completed these grades, and Level 2 was designed for students in Grades 10 to 12 and adults who have completed more than 9 years of schooling, but have not graduated from high school. The tests were designed primarily for educational and career counseling of students in Grades 7 to 12, but can also be used to assess abilities of less educated adults. The test also includes a Career Interest Inventory that can be used in conjunction with the aptitude tests, and a shortened version called the Differential Aptitude Tests for Personnel and Career Assessment (DAT for PCA) is packaged as a selection tool. The total time to administer the complete version of the DAT is slightly under 4 hours. The readability of the tests was assessed by The Psychological Corporation, and all vocabulary used in directions and content is at the fifth-grade reading level. The Psychological Corporation conducted a careful study of the tryout form of the test to make sure there was no racial test bias in items or scoring. Scoring of the test can be done by hand or computer, and there is a computerized version of the test.

The following separate tests are included:

Career Interest Inventory (30 minutes): Students indicate their level of interest in performing activities related to work and school.

Verbal Reasoning (40 items, 25 minutes): Items include analogies.

Numerical Reasoning (40 items, 30 minutes): Items include addition, subtraction, numeric sequences, fractions, multiplication, division, computing percentages, and basic algebra.

Abstract Reasoning (40 items, 20 minutes): Items assess logic, pattern or rule recognition, attention to detail, and abstract reasoning skills.

Perceptual Speed and Accuracy (2 parts, 100 items each part, 3 minutes each part): Test takers are asked to choose the letter/number combinations that are the same as the underlined combinations.

Mechanical Reasoning (60 items, 25 minutes): Test takers are presented with a picture of some mechanical principle and presented with a question.

Space Relations (50 items, 25 minutes): Items assess perceptual abilities, attention to detail, pattern recognition, and spatial relationships.

Spelling (40 items, 10 minutes): Test takers must determine which word is spelled incorrectly.

Language Usage (40 items, 15 minutes): Items include sentences with errors of grammar, capitalization, or punctuation that test takers are asked to identify.

—Jennifer Bragger

See also Aptitude Tests

Further Reading

Henly, S. J., Klebe, K. J., McBride, J. R., & Cudek, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement*, 13, 363–371.

Wang, L. (1995). Differential aptitude tests (DAT). *Measurement and Evaluation in Counseling and Development*, 28, 168–170.

Differential Aptitude Tests for Personnel and Career Assessment: <http://www.pantesting.com/products/PsychCorp/DAT.asp>

DIGGLE-KENWARD MODEL FOR DROPOUT

In medical research, studies are often designed in which specific parameters are measured repeatedly over time in the participating subjects. This allows for modeling the process of change within each subject separately, based on both subject-specific factors (such as gender) and experiment-specific factors (such as treatment). The analysis of such longitudinal data requires statistical models that take into account the association between the measurements within subjects. During the past decade, a lot of effort has been put into the search for flexible longitudinal models.

In practice, longitudinal studies often suffer from attrition (i.e., subjects dropping out earlier than scheduled) for reasons outside the control of the investigator. The resulting data are then unbalanced with unequal numbers of measures for each participant. Nowadays, several statistical packages can handle unbalanced longitudinal data. However, they yield valid inferences only under specific assumptions for the dropout process.

Generally, valid inferences can be obtained only by modeling the response measurements and the dropout process simultaneously. Making various assumptions about the dropout mechanism, a large variety of models for continuous as well as categorical outcomes have been proposed in the statistical literature. With the volume of literature on models for incomplete data increasing, there has been growing concern about the critical dependence of many of these models on the validity of the underlying assumptions. To compound the issue, the data often have very little to say about the correctness of such assumptions.

When referring to the missing-value, or nonresponse, process we will use the terminology of Little and Rubin. A nonresponse process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data, and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *nonrandom* (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are ignorable, whereas a nonrandom process is nonignorable. Ignorability implies that valid inferences about the measurement model parameters can be obtained by analyzing the observed data alone, obviating the need for formulation of a dropout model.

We will present one modeling framework that has been developed for incomplete longitudinal data of a continuous nature, proposed by Diggle and Kenward.

The model has been subject to criticism because it is rather vulnerable to the modeling assumptions made. These concerns will be discussed and a number of ways for dealing with it explored, with a prominent place given to sensitivity analysis.

The Diggle-Kenward Model for Dropout

We assume that for subject i in the study, $i = 1, \dots, N$, a sequence of measurements Y_{ij} is designed to be measured at time points t_{ij} , $j = 1, \dots, n_i$, resulting in a vector $Y_i = (Y_{i1}, \dots, Y_{in_i})'$ of measurements for each participant. If dropout occurs, Y_i is only partially observed. We denote the occasion at which dropout occurs by $D_i > 1$, and Y_i is split into the $(D_i - 1)$ -dimensional observed component Y_i^{obs} and the $(n_i - D_i + 1)$ -dimensional missing component Y_i^{mis} . In case of no dropout, we let $D_i = n_i + 1$, and Y_i equals Y_i^{obs} . The likelihood contribution of the i th subject, based on the observed data (y_i^{obs}, d_i) , is proportional to the marginal density function

$$\begin{aligned} f(y_i^{obs}, d_i | \theta, \psi) &= \int f(y_i, d_i | \theta, \psi) dy_i^{mis} \\ &= \int f(y_i | \theta) f(d_i | y_i, \psi) dy_i^{mis} \end{aligned} \quad (1)$$

in which a marginal model for Y_i is combined with a model for the dropout process, conditional on the response, and where θ and ψ are vectors of unknown parameters in the measurement model and dropout model, respectively.

Let $h_{ij} = (y_{i1}, \dots, y_{i,j-1})$ denote the observed history of subject i up to time $t_{i,j-1}$. The Diggle-Kenward model for the dropout process allows the conditional probability for dropout at occasion j , given that the subject was still observed at the previous occasion, to depend on the history h_{ij} and the possibly unobserved current outcome y_{ij} , but not on future outcomes y_{ik} , $k > j$. These conditional probabilities $P(D_i = j | D_i \geq j, h_{ij}, y_{ij}, \psi)$ can now be used to calculate the probability of dropout at each occasion:

$$\begin{aligned} P(D_i = j | y_i, \psi) &= P(D_i = j | h_{ij}, y_{ij}, \psi) \\ &= \begin{cases} P(D_i = j | D_i \geq j, h_{ij}, y_{ij}, \psi) & j = 2, \\ P(D_i = j | D_i \geq j, h_{ij}, y_{ij}, \psi) \\ \quad \times \prod_{k=2}^{j-1} [1 - P(D_i = k | D_i \geq k, h_{ik}, y_{ik}, \psi)] & j = 3, \dots, n_i, \\ \prod_{k=2}^{n_i} [1 - P(D_i = k | D_i \geq k, h_{ik}, y_{ik}, \psi)] & j = n_i + 1. \end{cases} \end{aligned}$$

Diggle and Kenward combine a multivariate normal model for the measurement process with a logistic regression model for the dropout process. More specifically, the measurement model assumes that the vector Y_i of repeated measurements for the i th subject satisfies the linear regression model $Y_i \sim N(X_i \beta, V_i)$, ($i = 1, \dots, N$). The matrix V_i can be left unstructured or is assumed to be of a specific form (e.g., resulting from a linear mixed model, a factor-analytic structure, or spatial covariance structure). The logistic dropout model is typically of the form

$$\begin{aligned} \text{logit}[P(D_i = j | D_i \geq j, h_{ij}, y_{ij}, \psi)] \\ = \psi_0 + \psi_1 y_{ij} + \psi_2 y_{i,j-1}. \end{aligned} \quad (2)$$

More general models can be constructed easily by including the complete history $h_{ij} = (y_{i1}, \dots, y_{i,j-1})$, as well as external covariates, in the above conditional dropout model. Note also that, strictly speaking, one could allow dropout at a specific occasion to be related to all future responses as well. However, this is rather counterintuitive in many cases. Moreover, including future outcomes seriously complicates the calculations because computation of the likelihood (Equation 1) then requires evaluation of a possibly high-dimensional integral. Note also that special cases of a model (Equation 2) are obtained from setting $\psi_1 = 0$ or $\psi_1 = \psi_2 = 0$, respectively. In the first case, dropout is no longer allowed to depend on the current measurement, implying random dropout (MAR). In the second case, dropout is independent of the outcome, which corresponds to completely random dropout (MCAR).

Diggle and Kenward obtained parameter and precision estimates by means of maximum likelihood. The likelihood involves marginalization over the unobserved outcomes Y_i^{mis} . Practically, this involves

relatively tedious and computationally demanding forms of numerical integration. This, combined with likelihood surfaces tending to be rather flat, makes the model difficult to use. These issues are related to the problems to be discussed next.

Remarks on Sensitivity Analysis and Other Models

Apart from the technical difficulties encountered during parameter estimation, there are further important issues surrounding MNAR based models. Even when the measurement model (e.g., the multivariate normal model) would be the choice of preference beyond any doubt to describe the measurement process *should the data be complete*, then the analysis of the actually observed, incomplete version is, in addition, subject to further untestable modeling assumptions.

When missingness is MAR, the problems are less complex because it has been shown that, in a likelihood or Bayesian framework, it is sufficient to analyze the observed data without explicitly modeling the dropout process. However, the very assumption of MAR is itself untestable. Therefore, ignoring MNAR models is as little an option as blindly shifting to one particular MNAR model. A sensible compromise between, on one hand, considering a single MNAR model or, on the other hand, excluding such models from consideration is to study the nature of such sensitivities and, building on this knowledge, formulate ways for conducting sensitivity analyses. Indeed, a strong conclusion, arising from most sensitivity analysis work, is that MNAR models have to be approached cautiously. This was made clear by several discussants to the original paper by Diggle and Kenward, particularly Laird, Little, and Rubin, respectively. An implication is that, for example, formal tests for the null hypothesis of MAR versus the alternative of MNAR should be approached with the utmost caution.

Verbeke, Lesaffre, and Spiessens have shown, in the context of an onychomycosis study, that excluding a small amount of measurement error drastically changes the likelihood ratio test statistics for the MAR null hypothesis. Kenward revisited the analysis of the

mastitis data performed by Diggle and Kenward. In this study, the milk yields of 107 cows were to be recorded during two consecutive years. Whereas data were complete in the first year, 27 animals were missing in the second year because they developed mastitis and their milk yield was no longer of use. In Diggle and Kenward's paper, there was strong evidence for MNAR, but Kenward showed that removing 2 out of 107 anomalous profiles completely removed this evidence. In addition, he showed that changing the conditional distribution of the Year 2 yield, given the Year 1 yield, from a normal distribution to a heavy-tailed t also led to the same result of no residual evidence for MNAR. This particular conditional distribution is of great importance, because a subject with missing data does not contribute to it and hence is a source of sensitivity issues. Once more, the conclusion is that fitting a MNAR model should be subject to careful scrutiny.

In addition to the instances described above, sensitivity to model assumptions has been reported for about two decades. In an attempt to formulate an answer to these concerns, a number of authors have proposed strategies to study sensitivity. We broadly distinguish between two types. A first family of approaches can be termed *substantive driven* in the sense that the approaches start from particularities of the problem at hand. Kenward's approach falls within this category. Arguably, such approaches are extremely useful, both in their own right and as a preamble to using the second family, where what could be termed *general purpose* tools are used.

Broadly, we could define a sensitivity analysis as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools (such as diagnostic measures). This rather loose and very general definition encompasses a wide variety of useful approaches. The simplest procedure is to fit a selected number of (MNAR) models that are all deemed plausible or one in which a preferred (primary) analysis is supplemented with a number of variations. The extent to which conclusions (inferences) are stable across such ranges provides an indication about the belief that can be put into them. Variations to a basic model

can be constructed in different ways. The most obvious strategy is to consider various dependencies of the missing data process on the outcomes and/or on covariates. Alternatively, the distributional assumptions of the models can be changed.

Several authors have proposed the use of global and local influence tools. Molenberghs, Verbeke, Thijs, Lesaffre, and Kenward revisited the mastitis example. They were able to identify the same two cows found by Kenward, in addition to another one. Thus, an important question is, What exactly are the sources causing an MNAR model to provide evidence for MNAR against MAR? There is evidence to believe that a multitude of outlying aspects, but not necessarily the (outlying) nature of the missingness mechanism in one or a few subjects, is responsible for an apparent MNAR mechanism. The consequence of this is that local influence should be applied and interpreted with due caution.

Of course, the above discussion is not limited to the Diggle-Kenward model. A variety of other models have been proposed for incomplete longitudinal data. First, the model has been formulated within the selection model framework, in which the joint distribution of the outcome and dropout processes is factorized as the marginal distribution of the outcomes $f(y_i | \theta)$ and the conditional distribution of the dropout process, given the outcomes $f(d_i | y_i, \psi)$. Within this framework, models have been proposed for non-monotone missingness as well, and furthermore, a number of proposals have been made for non-Gaussian outcomes. Apart from the selection model framework, so-called pattern-mixture models have gained popularity, where the reverse factorization is applied with factors $f(y_i | d_i, \theta)$ and $f(d_i | \psi)$. Also within this framework, both models and sensitivity analysis tools for them have been formulated. A third framework consists of so-called shared parameter models, where random effects are employed to describe the relationship between the measurement and dropout processes.

—Geert Verbeke and Geert Molenberghs

See also Longitudinal/Repeated Measures Data; Missing Data Method; Mixed Models; Repeated Measures Analysis of Variance

Further Reading

- De Gruttola, V., & Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50, 1003–1014.
- Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49–93.
- Diggle, P. J., Heagerty, P. J., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford, UK: Clarendon.
- Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine*, 17, 2723–2732.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Molenberghs, G., & Verbeke, G. (2005). *Discrete longitudinal data*. New York: Springer.
- Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Mastitis in dairy cattle: Local influence to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, 37, 93–113.
- Verbeke, G., Lesaffre, E., & Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, 35, 419–434.

DIMENSION REDUCTION

Dimension reduction is a collection of statistical methodologies that reduces the dimension of the data while still preserving relevant information. High-dimensional data are very common in government agencies, academia, and industrials. However, the high dimension and large volume of data bring up at least two issues, among many others. One is to overcome the curse of dimensionality, which states that high-dimensional spaces are inherently sparse even with large number of observations. The other is how to present the information within data parsimoniously. Dimension reduction techniques address these issues to varying extents by reducing the set of variables to a smaller set of either the original variables or new variables, where the new variables are linear combinations or even nonlinear functions of the original ones. When the new dimension is relatively small, data visualization becomes possible, which often assists data modeling substantially.

Dimension Reduction Methodologies

Based on whether a response is specified or not, dimension reduction techniques generally can be divided into two major categories: supervised dimension reduction and unsupervised dimension reduction.

Unsupervised Dimension Reduction

Unsupervised dimension reduction treats all variables equally without specifying a response. The analysis usually has a natural definition about the information of interest. Unsupervised dimension reduction methods find a new set of a smaller number of variables that either provides a simpler presentation or discovers intrinsic structure in the data while retaining most of the important information. Listed below are only a few of the most widely used techniques.

Principal component analysis (PCA) finds a few orthogonal linear combinations of the original variables with the largest variances; these linear combinations are the principal components that would be retained for subsequent analyses. In PCA, the information is the variation within the data. Usually, principal components are sorted in descending order according to their variations. The number of principal components that should be included in the analysis depends on how much variation should be preserved.

Factor analysis assumes that a set of variables establishes the relationships among themselves through a smaller set of common factors. It estimates the common factors with assumptions about the variance-covariance structure.

Canonical correlation analysis identifies and measures the association between two sets of random variables. Often, it finds one linear combination of variables for each set, where these two new variables have the largest correlation.

Correspondence analysis is a graphical tool for an exploratory data analysis of a contingency table. It projects the rows and columns as points into a plot, where rows (columns) have a similar profile if their corresponding points are close together.

Projection pursuit defines a projection index that measures the “interestingness” of a direction. Then, it searches for the direction maximizing the index.

Multidimensional scaling finds a projection of the data into a smaller dimensional space so that the distances among the points in the new space reflect the proximities in the original data.

Supervised Dimension Reduction

Supervised dimension reduction techniques generally are applied in regression. A response Y is specified that can be one random variable, one random vector, or even a curve. The predictor vector X is p -dimensional. The object of interest is the relation between the response and the predictors, which is often summarized as $Y = f(X, \epsilon)$, where ϵ denotes the error term. Some specific structures are imposed to facilitate the estimation of the function. Dimension reduction is a crucial part of the modeling process. For example, ordinary least squares regression can be considered as a special case of dimension reduction in regression.

To reduce the dimension in the predictor space, variable selection techniques select a small set of variables that is necessary instead of the whole p predictors. Single-index and multi-index models focus on only one or a small number of linear combinations of predictors. For example, a multi-index model assumes $Y = f(\beta_1^T X, \beta_2^T X, \dots, \beta_k^T X, \epsilon)$, where $\beta_i^T X$, $i = 1, 2, \dots, k$ are linear combinations of X and ϵ is an error term.

In addition to reducing the dimension of the predictor space, we also can apply dimension reduction in the functional space. For example, a generalized additive model assumes $Y = \sum_{i=1}^p f_i(x_i) + \epsilon$, where the additive nature of predictors' contributions to the model dramatically reduces the functional space considered for the regression.

Sufficient dimension reduction (SDR) in regression has generated considerable interest in the past decade. The basic idea is to replace the predictor vector with its projection onto a subspace of the predictor space without loss of information on the conditional distribution of $Y|X$. Specifically, suppose the response Y is independent of X given the values of d linear combinations of predictors $(\beta_1^T X, \beta_2^T X, \dots, \beta_d^T X)$. Thus, these d new variables carry all the information that X has about Y . The subspace spanned by the β_i s is called the SDR subspace. One advantage

of working in SDR is that no prespecified model for $Y|X$ is required. Many methods have been proposed to estimate the SDR subspace.

Sliced inverse regression (SIR) is one of the most widely used methods. Without loss of generality, we assume X is a standardized predictor vector with a mean of zero and a covariance matrix as an identity matrix. Under mild conditions, the inverse conditional means of X given the response Y belong to the SDR subspace. When Y is discrete, it is easy to calculate the sample version of inverse conditional means. If Y is continuous, we only need to discretize Y by slicing on the range of the response. Suppose we have h slices. Let $\bar{X}_s$ denote the sample average of X within the s th slice, $s = 1, 2, \dots, h$. Construct a SIR kernel matrix $M_{SIR} = \sum_{s=1}^h f_s \bar{X}_s \bar{X}_s^T$, where f_s is the proportion of the observations falling in the s th slice. If we determine that the dimension of the SDR subspace is d , then the d eigenvectors of M_{SIR} that correspond to the d largest eigenvalues constitute a basis for the SDR subspace.

Sliced average variance estimation (SAVE) is another important SDR method. Under the same setting as SIR, SAVE constructs a kernel matrix using the inverse conditional variance of X given Y : $M_{SAVE} = \sum_{s=1}^h f_s (I - \Omega_s)^2$, where Ω_s is the sample covariance matrix of X within the s th slice. As with SIR, the first few eigenvectors of M_{SAVE} serve as an estimated basis of the SDR subspace.

Principal Hessian directions (pHd) is an SDR method that does not require slicing. Its kernel matrix is constructed as $M_{pHd} = 1/n \sum_{i=1}^n (y_i - \bar{y}) X_i X_i^T$, where $\bar{y}$ is the sample average of the response and n is the number of observations. The estimated basis of the SDR subspace is the eigenvectors of M_{pHd} that correspond to eigenvalues with the largest absolute values.

Minimum average variance estimation (MAVE), which is one of the more recent developments, has virtually no assumptions on X but is computationally more intensive than SIR or SAVE. MAVE essentially is a local linear smoother with weights determined by some kernel functions.

Case Study

We consider a data set of 200 Swiss banknotes, among which half are genuine. The predictors are six

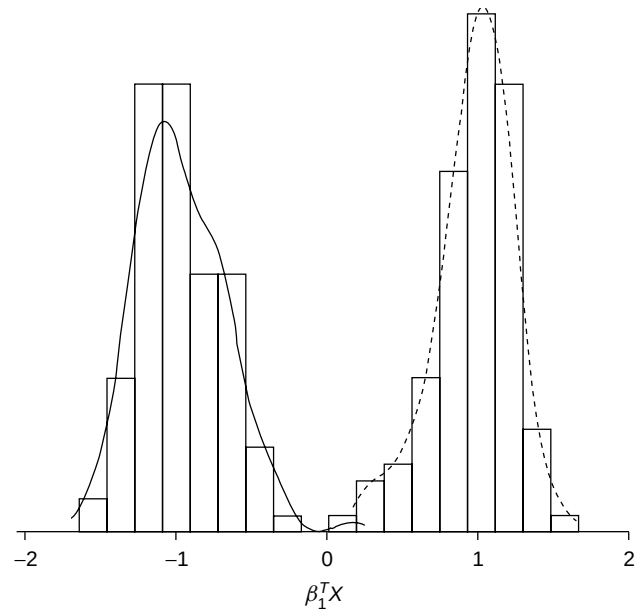


Figure 1 SIR Result of Swiss Banknotes Data

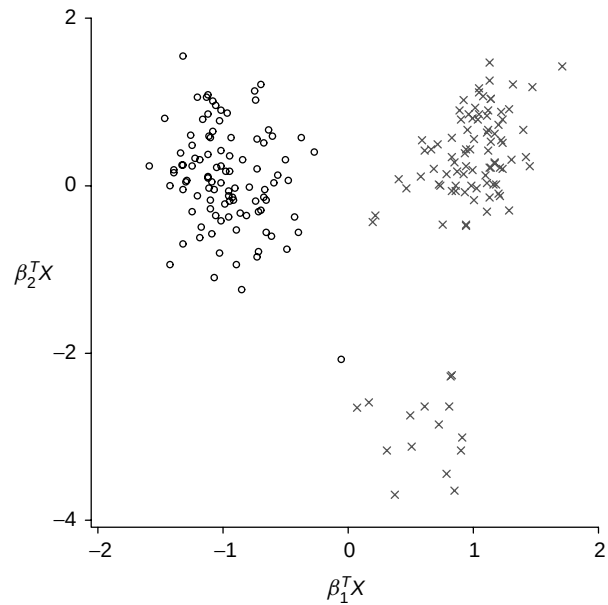


Figure 2 SAVE Result for Swiss Banknotes Data

measurements: the widths of the top margin, the bottom margin, the left edge, the right edge, the length of the bill, and the length of the image diagonal. The response is the status of the bill, which equals 1 if the bill is counterfeit and 0 otherwise. In this case, the response is discrete with two slices.

SIR can detect only one direction β_1 in the SDR subspace because there are only two slices. Figure 1 plots the histogram of the linear combination of the predictors $\beta_1^T X$, where smoothing density curves for both groups have been superimposed for visual enhancement. The left solid curve is for genuine banknotes. The direction that SIR detected separates these two groups very well.

For the same data, SAVE can detect more than one direction. Figure 2 plots the first two directions from SAVE, where circles denote genuine notes and crosses denote counterfeit notes. The first SAVE direction is almost identical to the SIR direction. The second SAVE direction brought up an interesting pattern in the data, which may suggest that there are two sources of counterfeit notes. It is also possible that this pattern is just a result of measurement errors.

—Liqiang Ni

Further Reading

- Cook, R. D. (2000, June). *Using Arc for dimension reduction and graphical exploration in regression*. Retrieved from <http://www.stat.umn.edu/arc/InvReg/DimRed.pdf>
- Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Society*, 100, 410–428.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice Hall.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Society*, 86, 316–342.

DISCRIMINANT ANALYSIS

The need for classification arises in most scientific pursuits. Typically, there is interest in classifying an entity, say, an individual or object, on the basis of some characteristics (feature variables) measured on the entity. This classification is usually undertaken in the context where there is a finite number, say, g , of predefined distinct populations, categories, classes, or groups, and the entity to be classified is assumed to belong to one (and only one) of these g possible groups. In order to assist with the construction of a classification rule or classifier for this purpose, there

are usually available so-called training data from each group; that is, these training data comprise the features measured on some entities that are classified with respect to the g underlying groups. In statistical terminology, this classification process is referred to as *discriminant analysis*, whereas in pattern recognition and machine learning, it is referred to as *supervised learning* or *learning with a teacher*.

We let $G_1, \dots, G_g$ denote the g possible groups, and we suppose that a (feature) vector x containing p variables can be measured on the entity. The group membership of the entity is denoted by the categorical variable z , where $z = i$ implies that the entity belongs to G_i ($i = 1, \dots, g$). The problem is to estimate or predict z solely on the basis of x and the associated training data. An example in which an outright assignment is required concerns the rejection or acceptance of loan applicants by a financial institution. For this decision problem, there are two groups: G_1 refers to applicants who will service their loans satisfactorily, and G_2 refers to those who will not. The feature vector x for an applicant contains information such as age, income, and marital status. A rule based on x for allocating an applicant to either G_1 or G_2 (that is, either accepting or rejecting the loan application) can be formed from an analysis of the feature vectors of past applicants from each of the two groups. In some applications, no assignment of the entity to one of the possible groups is intended. Rather, the problem is to draw inferences about the relationship between z and the feature variables in x . An experiment might be designed with the specific aim to provide insight into the predictive structure of the feature variables. For example, a political scientist may wish to determine the socioeconomic factors that have the most influence on the voting patterns of a population of voters.

Allocation Rules

Let $r(x)$ denote an allocation or discriminant rule, where $r(x) = i$ implies that an entity with feature vector x is to be assigned to the i th group G_i . The allocation rates associated with this rule $r(x)$ are denoted by $e_{ij}(r)$, where $e_{ij}(r)$ is the probability that a randomly chosen entity from G_i is allocated to G_j ($i, j = 1, \dots, g$).

For a diagnostic test using the rule $r(x)$ in the context where G_1 denotes the absence of a disease or condition and G_2 its presence, the error rate $e_{12}(r)$ corresponds to the probability of a false positive, whereas $e_{21}(r)$ is the probability of a false negative. The correct allocation rates $e_{22}(r)$ and $e_{11}(r)$ are known as the sensitivity and specificity, respectively, of the diagnostic test.

Decision theory provides a convenient framework for the construction of discriminant rules. More specifically, let π_i denote the prior probability that the entity comes from the i th group G_i , and let $f_i(x)$ denote the probability density function of the feature vector X in G_i . This is assuming that the variables in the feature vector x are of the continuous type. The unconditional density of X , $f(x)$, is therefore given by the mixture density, which is the sum of the group-conditional densities $f_i(x)$ weighted by their prior probabilities π_i . If they are discrete variables, then $f_i(x)$ can be viewed as the probability function. With this notation, the posterior probability that an entity belongs to G_i can be expressed via Bayes' theorem as

$$\tau_i(x) = \pi_i f_i(x) / f(x) \quad (i=1, \dots, g) \quad (1)$$

An optimal rule of allocation can be formed by assigning an entity with feature vector x to that group to which the entity has the greatest posterior probability of belonging. This rule is optimal in the sense of minimizing the overall correct allocation rate. It can be viewed also as optimal in a decision-theoretic framework of minimizing the so-called risk of allocation under certain assumptions on the loss function.

Sample-Based Allocation Rules

In practice, the group-conditional densities $f_i(x)$ are usually unknown. A basic assumption in discriminant analysis is that in order to estimate the unknown group-conditional densities, there are entities of known origin on which the feature vector X has been recorded for each. These data are referred to in the literature as *initial*, *reference*, *design*, *training*, or *learning data*.

The initial approach to the problem of forming a sample discriminant rule, and indeed to discriminant

analysis in its modern guise, was developed by Fisher. In the context of $g = 2$ groups, he proposed that an entity with feature vector x be assigned on the basis of the linear discriminant function $a^T x$, where a maximizes an index of separation between the two groups. The index was defined to be the magnitude of the difference between the group sample means of $a^T x$ normalized by the pooled sample estimate of its assumed common variance within a group.

We let $r(x;t)$ denote a sample-based allocation rule formed from the training data t . An obvious way of forming $r(x;t)$ is to take it to be an estimated version of Bayes' rule, where the posterior probabilities of group membership $\tau_i(x)$ are replaced by some estimates $\hat{\tau}_i(x;t)$ formed from the training data t . A common approach, referred to as the sampling approach, is to formulate the $\tau_i(x)$ through the group-conditional densities $f_i(x)$. With the fully parametric approach to this problem, the group-conditional densities $f_i(x)$ are assumed to have specified functional forms except for a finite number of parameters to be estimated. A commonly used parametric family for the $f_i(x)$ for continuous feature data is the normal with either a linear or quadratic rule in x being obtained depending on whether the group covariance matrices are taken to be equal or unequal.

There is also the direct approach (the diagnostic paradigm) to the estimation of the $\tau_i(x)$, using either nonparametric estimates, as with nearest neighbor methods, or parametric estimates via the logistic model. With the latter approach, it is the ratios of the group-conditional densities that are being modeled. The fundamental assumption of the logistic approach is that these ratios are linear, which is equivalent to taking the log (posterior) odds to be linear. The linearity here is not necessarily in the basic variables; transforms of these may be taken. Another method for the direct modeling of the group posterior probabilities is to use neural networks, which have been coming under increasing attention by statisticians.

To illustrate the construction of sample-based allocation rules, we consider the construction of a discriminant rule for the problem of assessing loan applications, as mentioned earlier. Here, the applicant for a loan is a company, and the feature vector

contains information on its key financial characteristics. For the purposes of our illustration, we consider the case where there are $p = 2$ feature variables, which are the first two principal components of three financial ratios concerning the debt, liquid assets, and working capital of a company. In Figure 1, we have plotted the boundary

$$\hat{\tau}_1(x, t) = \hat{\tau}_2(x, t)$$

for the sample version of the Bayes' rule (the sample linear discriminant rule) formed under the assumption that the feature vector has a bivariate normal distribution with a common group covariance matrix in each of the two groups G_1 and G_2 , corresponding to good (satisfactorily serviced) and bad (unsatisfactorily serviced) loans. It can be seen that the overall error rate of this rule reapplied to the training data (its apparent error rate) is not as low as that of the quadratic rule formed without the restriction of a common group-conditional covariance matrix for the feature vector X . With its curved boundary, the quadratic rule misallocates only four (all bad) loans. One should be mindful, however, that the apparent error rate provides an optimistic assessment of the accuracy of a discriminant rule when it is applied to data not in the training set. We have also plotted in Figure 1 the boundary of the rule obtained by modeling the distribution of the feature vector X by a two-component normal mixture, which results in the misallocation of one bad loan at the expense of misallocating two good loans. The latter rule is an example of a more flexible approach to classification than that based on the assumption of normality for each of the group-conditional distributions, as to be considered now.

Flexible Discriminant Rules

A common nonparametric approach to discriminant analysis uses the kernel method to estimate the group-conditional densities $f_i(x)$ in forming an estimate of the Bayes' rule. More recently, use has been made of finite mixture models, mainly normal mixtures, to provide flexible rules of discrimination. Mixture models, which provide an extremely flexible way

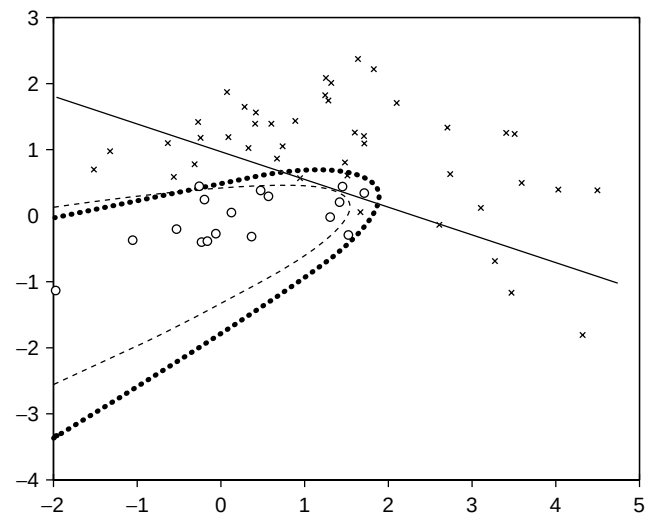


Figure 1 Boundaries of Linear Rule (solid line), Quadratic Rule (dashed line), and Mixture Model-Based Rule (dotted line)

Note: Crosses and circles denote good and bad loans, respectively.

of modeling a density function, can be fitted in a straightforward manner via the Expectation-Maximization algorithm. Among other work on flexible discrimination, there is the flexible discriminant analysis (FDA) approach based on nonparametric regression. The generic version of FDA based on smoothing splines proceeds by expanding the predictors in a large (adaptively selected) basis set, and then performing a penalized discriminant analysis in the enlarged space using a linear discriminant or a normal mixture model-based rule. The class of nonlinear regression methods that can be used includes additive models, the multivariate adaptive regression spline (MARS) model, projection pursuit regression, and neural networks. In machine learning, there has been increasing attention in the case of two groups given to nonlinear rules based on the foundations of support vector machines. With this approach, the initial feature space is mapped into a higher dimensional space by choosing a nonlinear mapping and then choosing an optimal separating hyperplane in the enlarged feature space.

A rather different approach to the allocation problem as considered up to now, is to portray the rule in

terms of a binary tree. The tree provides a hierarchical representation of the feature space. An allocation is effected by proceeding down the appropriate branches of the tree. The Classification and Regression Tree (CART) methodology of Breiman et al. has contributed significantly to the growing popularity of tree classifiers. In the context of tree classifiers in particular, there has been growing interest in the use of boosting, which is one of the most important recent developments in classification methodology. Algorithms such as the Adaboost algorithm of Freund and Schapire and the bagging algorithm often can improve performance of unstable classifiers like trees or neural nets by applying them sequentially to reweighted versions of the training data and taking a weighted majority vote of the sequence of classifiers so formed. The test error of the weighted classifier usually does not increase as its size increases, and often is observed to decrease even after the training error reaches zero.

—Geoffrey McLachlan

See also Discriminant Correspondence Analysis

Further Reading

- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
- Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine learning: Proceedings of the Thirteenth International Conference* (L. Saitta, ed., pp. 148–156). San Francisco: Morgan Kaufmann.
- Hand, D. J. (1997). *Construction and assessment of classification rules*. New York: Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

DISCRIMINANT CORRESPONDENCE ANALYSIS

As the name indicates, discriminant correspondence analysis (DCA) is an extension of discriminant analysis (DA) and correspondence analysis (CA). Like discriminant analysis, the goal of DCA is to categorize observations in predefined groups, and like correspondence analysis, it is used with nominal variables.

The main idea behind DCA is to represent each group by the sum of its observations and to perform a simple CA on the groups by variables matrix. The original observations are then projected as supplementary elements, and each observation is assigned to the closest group. The comparison between the a priori and a posteriori classifications can be used to assess the quality of the discrimination. A similar procedure can be used to assign new observations to categories. The stability of the analysis can be evaluated using cross-validation techniques such as jackknifing or bootstrapping.

An Example

It is commonly thought that the taste of wines depends upon their origin. As an illustration, we have sampled 12 wines coming from three different origins (four wines per origin) and asked a professional taster (unaware of the origin of the wines) to rate these wines on five scales. The scores of the taster were then transformed into binary codes to form an indicator matrix (as in multiple correspondence analysis). For example, a score of 2 on the “Fruity” scale would be coded by the following pattern of three binary values: 010. An additional unknown wine was also evaluated by the taster with the goal of predicting its origin from the ratings. The data are given in Table 1.

Notations

There are K groups, with each group comprising I_k observations, and the sum of the I_k s is equal to I , which is the total number of observations. For convenience, we assume that the observations constitute the rows of the data matrix, and that the variables are the

Table 1 Data for the Three Region Wines Example

Wine	Region	Woody			Fruity			Sweet			Alcohol			Hedonic			
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4
1	1 Loire	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0
2	1 Loire	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0
3	1 Loire	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0
4	1 Loire	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1
Σ	1 Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1
5	2 Rhône	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0
6	2 Rhône	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0
7	2 Rhône	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0
8	2 Rhône	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1
Σ	2 Rhône	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1
9	3 Beaujolais	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	0
10	3 Beaujolais	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0
11	3 Beaujolais	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1
12	3 Beaujolais	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0
Σ	3 Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1
W?	?	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0

Notes: Twelve wines from three different regions are rated on five descriptors. A value of 1 indicates that the wine possesses the given value of the variable. The wine W? is an unknown wine treated as a supplementary observation.

columns. There are J variables. The $I \times J$ data matrix is denoted $\mathbf{X}$. The *indicator* matrix is an $I \times K$ matrix denoted $\mathbf{Y}$ in which a value of 1 indicates that the row belongs to the group represented by the column, and a value of 0 indicates that it does not. The $K \times J$ matrix, denoted $\mathbf{N}$, is called the “group matrix,” and it stores the total of the variables for each category. For our example, we find that

$$\mathbf{N} = \mathbf{Y}^T \mathbf{X}$$

$$= \begin{bmatrix} 3 & 1 & 0 & 0 & 1 & 3 & 0 & 2 & 2 & 2 & 2 & 0 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 2 & 1 & 2 & 1 & 1 & 0 & 1 & 3 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 3 & 1 & 0 & 1 & 1 & 2 & 3 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}. \quad (1)$$

Performing CA on the group matrix $\mathbf{N}$ provides two sets of factor scores—one for the groups (denoted $\mathbf{F}$) and one for the variables (denoted $\mathbf{G}$). These factor scores are, in general, scaled such that their variance is equal to the eigenvalue associated with the factor.

The grand total of the table is noted N , and the first step of the analysis is to compute the probability matrix $\mathbf{Z} = N^{-1}\mathbf{N}$. We denote $\mathbf{r}$ the vector of the row totals of $\mathbf{Z}$ (i.e., $\mathbf{r} = \mathbf{Z}\mathbf{1}$, with $\mathbf{1}$ being a conformable vector of 1s); $\mathbf{c}$ the vector of the column totals; and $\mathbf{D}_c = \text{diag}\{\mathbf{c}\}$, $\mathbf{D}_r = \text{diag}\{\mathbf{r}\}$. The factor scores are obtained from the following singular value decomposition:

$$\mathbf{D}_r^{-\frac{1}{2}}(\mathbf{Z} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^T. \quad (2)$$

($\mathbf{\Lambda}$ is the diagonal matrix of the *singular* values, and $\mathbf{\Lambda} = \mathbf{\Delta}^2$ is the matrix of the *eigenvalues*.) The row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{P}\mathbf{\Lambda} \quad \text{and} \quad \mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{Q}\mathbf{\Lambda}. \quad (3)$$

The squared (χ^2) distances from the rows and columns to their respective barycenters are obtained as

$$\mathbf{d}_r = \text{diag}\{\mathbf{F}\mathbf{F}^T\} \quad \text{and} \quad \mathbf{d}_c = \text{diag}\{\mathbf{G}\mathbf{G}^T\}. \quad (4)$$

The squared cosines between row i and factor ℓ and column j and factor ℓ are obtained respectively as

$$o_{i,\ell} = \frac{f_{i,\ell}^2}{d_{r,i}^2} \quad \text{and} \quad o_{j,\ell} = \frac{g_{j,\ell}^2}{d_{c,j}^2} \quad (5)$$

(with $d_{r,i}^2$ and $d_{c,j}^2$ being respectively the i th element of $\mathbf{d}_r$ and the j th element of $\mathbf{d}_c$). Squared cosines help locate the factors important for a given observation. The contributions of row i to factor ℓ and of column j to factor ℓ are obtained respectively as

$$t_{i,\ell} = \frac{f_{i,\ell}^2}{\lambda_\ell} \quad \text{and} \quad t_{j,\ell} = \frac{g_{j,\ell}^2}{\lambda_\ell}. \quad (6)$$

Contributions help locate the observations important for a given factor.

Supplementary or illustrative elements can be projected onto the factors using the so-called transition formula. Specifically, let $\mathbf{i}_{\text{sup}}^T$ be an illustrative row and $\mathbf{j}_{\text{sup}}$ be an illustrative column to be projected. Their coordinates $\mathbf{f}_{\text{sup}}$ and $\mathbf{g}_{\text{sup}}$ are obtained as

$$\mathbf{f}_{\text{sup}} = (\mathbf{i}_{\text{sup}}^T \mathbf{1}) \mathbf{i}_{\text{sup}}^T \mathbf{G} \mathbf{\Delta}^{-1} \quad \text{and} \quad \mathbf{g}_{\text{sup}} = (\mathbf{j}_{\text{sup}}^T \mathbf{1}) \mathbf{j}_{\text{sup}}^T \mathbf{F} \mathbf{\Delta}^{-1}. \quad (7)$$

After the analysis has been performed on the groups, the original observations are projected as supplementary elements and their factor scores are stored in a matrix denoted $\mathbf{F}_{\text{sup}}$. To compute these scores, first compute the matrix of row profiles $\mathbf{R} = (\text{diag}\{\mathbf{X}\mathbf{1}\})^{-1} \mathbf{X}$ and then apply Equation 7 to obtain

$$\mathbf{F}_{\text{sup}} = \mathbf{R} \mathbf{G} \mathbf{\Delta}^{-1}. \quad (8)$$

The Euclidean distance between the observations and the groups computed from the factor scores is equal to the χ^2 -distance between their row profiles. The $I \times K$ distance matrix between observations and groups is computed as

$$\mathbf{D} = \mathbf{s}_{\text{sup}} \mathbf{1}^T + \mathbf{1} \mathbf{s}^T - 2 \mathbf{F}_{\text{sup}} \mathbf{F}^T \quad \text{with} \quad (9)$$

$$\mathbf{s}_{\text{sup}} = \text{diag}\{\mathbf{F}_{\text{sup}} \mathbf{F}_{\text{sup}}^T\} \quad \text{and} \quad \mathbf{s} = \text{diag}\{\mathbf{F} \mathbf{F}^T\}.$$

Each observation is then assigned to the closest group.

Model Evaluation

The quality of the discrimination can be evaluated as a fixed-effect model or as a random-effect model. For the fixed-effect model, the correct classifications are compared to the assignments obtained from Equation 9. The fixed-effect model evaluates the quality of the classification on the sample used to build the model.

The random-effect model evaluates the quality of the classification on new observations. Typically, this step is performed using cross-validation techniques such as jackknifing or bootstrapping.

Results

Tables 2 and 3 give the results of the analysis and Figure 1 displays them. The fixed-effect quality of the model is evaluated by the following confusion matrix:

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 1 & 4 \end{bmatrix}. \quad (10)$$

In this matrix, the rows are the assigned groups and the columns are the real groups. For example, out of five wines assigned to the wine region Beaujolais (Group 3), one wine was, in fact, from the Rhône region (Group 2), and four wines were from Beaujolais. The overall quality can be computed from the diagonal of the matrix. Here, we find that 11 (4 + 3 + 4) wines out of 12 were classified correctly.

A jackknife procedure was used in order to evaluate the generalization capacity of the analysis to new wines (i.e., this corresponds to a random-effect analysis). Each wine was, in turn, taken out of the sample, a DCA was performed on the remaining sample of 11 wines, and the wine taken out was assigned to the closest group. This gave the following confusion matrix:

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \quad (11)$$

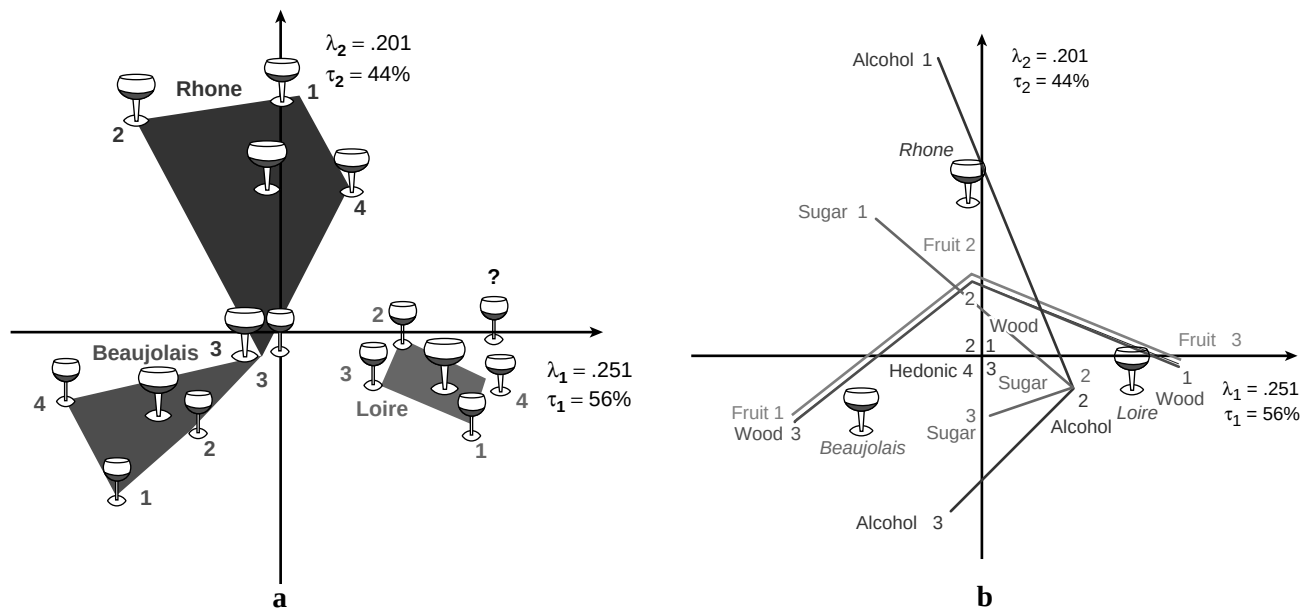


Figure 1 Discriminant Correspondence Analysis

Notes: Projections on the first 2 dimensions. (a) The *I* set: rows (i.e., wines). The wines are projected as supplementary elements, and *Wine?* is an unknown wine. (b) The *J* set: columns (i.e., descriptors). The wine categories have also been projected for ease of interpretation. Both figures have the same scale (some projection points have been slightly moved to increase readability). (Projections from Tables 2 and 3).

As expected, the performance of the model as a random effect is less impressive than as a fixed-effect model. Now, only 6 (2 + 2 + 2) wines out of 12 are classified correctly.

The differences between the fixed- and the random-effect models are illustrated in Figure 2, where the jackknifed wines have been projected onto the fixed-effect solution (using metric multidimensional

Table 2 Factor Scores, Squared Cosines, and Contributions for the Variables (*J* set)

			Woody			Fruity			Sweet			Alcohol			Hedonic			
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4
Axis	λ	%	Factor Scores															
1	.251	55	.93	-.05	-.88	-.88	-.05	.93	-.51	.33	.04	-.14	.33	-.20	0	0	0	0
2	.201	44	-.04	.35	-.31	-.31	.35	-.04	.64	-.13	-.28	-.74	-.13	1.40	0	0	0	0
Axis	Squared Cosines																	
1			.998	.021	.892	.892	.021	.998	.384	.864	.021	.035	.864	.021	0	0	0	0
2			.002	.979	.108	.108	.979	.002	.616	.137	.979	.965	.137	.979	0	0	0	0
Axis	Contributions																	
1			.231	.001	.207	.207	.001	.231	.051	.029	.001	.007	.029	.008	0	0	0	0
2			.0006	.0405	.0313	.0313	.0405	.0006	.1019	.0056	.0324	.2235	.0056	.4860	0	0	0	0

Note: Contributions corresponding to negative scores are in italic.

Table 3 Factor Scores, Squared Cosines, and Contributions for the Regions

Axis	λ	%	Loire				Rhône				Beaujolais							
			Wines				Wines				Wines							
			Region		Region		Region		Region		Region		Region					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	W?		
Factor Scores																		
1	251	55	0.66	0.82	0.50	0.43	0.89	-0.10	0.07	-0.66	-0.11	0.29	-0.56	-0.74	-0.41	-0.11	-0.96	1.01
2	201	44	-0.23	-0.42	-0.05	-0.25	-0.22	0.63	1.05	0.93	-0.10	0.64	-0.39	-0.73	-0.43	-0.10	-0.32	-0.15
Squared Cosines																		
1			.89	.79	.99	.75	.94	.03	.00	.33	.56	.17	.67	.51	.47	.56	.90	.98
2			.11	.21	.01	.25	.06	.97	1.00	.67	.44	.83	.33	.49	.53	.44	.10	.02
Contributions																		
1			.58	.	.	.	.	.01	.	.	.	.	.41	.	.	.	.	.
2			.09	.	.	.	.	.65	.	.	.	.	.26	.	.	.	.	.

Note: The supplementary rows are the wines from the region and the mysterious wine (W?). Contributions corresponding to negative scores are in italic.

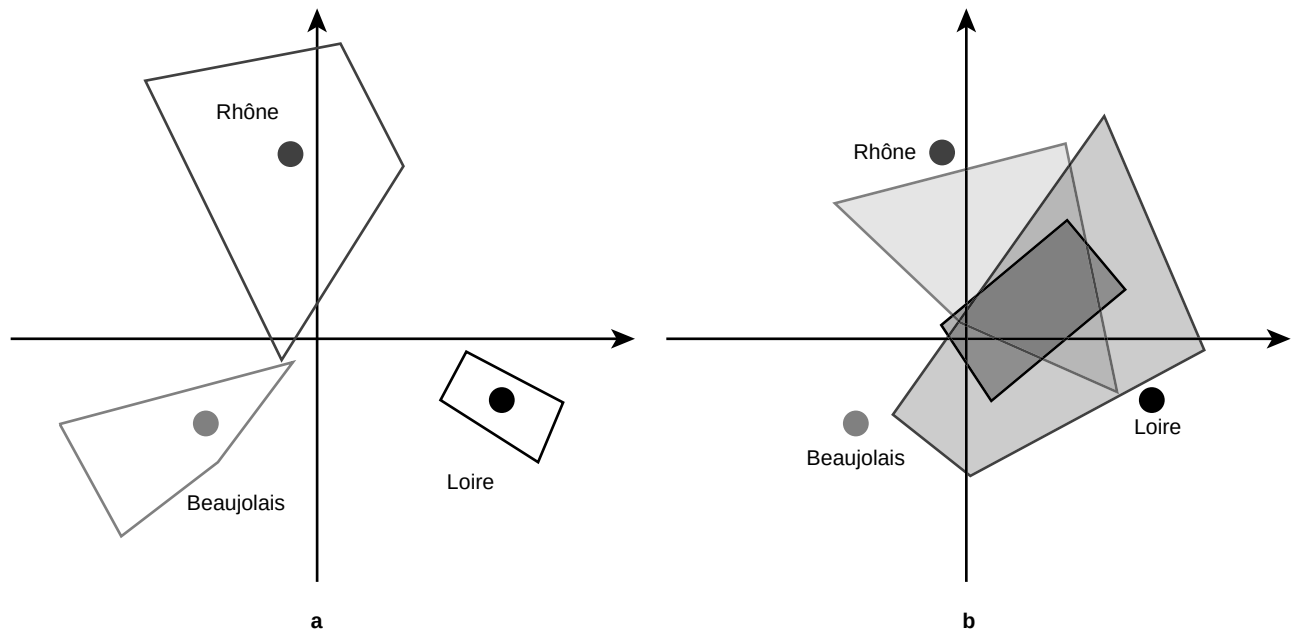


Figure 2 Discriminant Correspondence Analysis

Note: Projections on the first 2 dimensions. (a) Fixed effect model. The three regions and the convex envelop for the wines (b) Random effect model. The jackknifed wines have been projected back onto the fixed effect solution. The convex envelop shows that the random effect categories have a larger variability and have moved.

scaling). The quality of the model can be evaluated by drawing the convex envelop of each category. For the fixed-effect model, the centers of gravity of the convex envelops are the categories, and this illustrates that DCA is a least square estimation technique. For the random-effect model, the degradation of performance is due to a larger variance (the areas of the convex envelops are larger) and to a rotation of the envelops (the convex envelops are no longer centered on the category centers of gravity).

—Hervé Abdi

See also Centroid; Correspondence Analysis; Discriminant Analysis; Distance; Metric Multidimensional Scaling; Multiple Correspondence Analysis

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Clausen, S. E. (1998). *Applied correspondence analysis*. Thousand Oaks, CA: Sage.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.

Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Newbury Park, CA: Sage.

DISSIMILARITY COEFFICIENT

A dissimilarity coefficient is a function that measures the difference between two objects. It is defined from a set $E \times E$ (e.g., $\mathbb{R} \times \mathbb{R}$, $\mathbb{R}^2 \times \mathbb{R}^2$, $\mathbb{R}^n \times \mathbb{R}^n$) to the non-negative real numbers $\mathbb{R}^+$. Let g be a dissimilarity coefficient. Let x and y be two elements from E , and g verifies the following properties:

$$g(x,x) = 0 \quad (\text{C1}),$$

$$g(x,y) = g(y,x) \quad (\text{C2: symmetry}),$$

$$g(x,y) \geq 0 \quad (\text{C3: positivity}).$$

The function g is said to be a pseudo-metric if and only if g verifies C1, C2, C3, and the following property. Let z be another element from E ,

$$g(x,y) + g(y,z) \geq g(x,z) \quad (\text{C4: triangle inequality}).$$

Furthermore, the function g is said to be a metric if and only if g verifies C1, C2, C3, C4, and the following additional property:

$$g(x, y) = 0 \rightarrow x = y \quad (\text{C5}).$$

The value taken by g for two elements x and y is called “dissimilarity” if g is simply a dissimilarity coefficient; “semi-distance” if g is, in addition, a pseudo-metric; and “distance” if g is a metric.

The application of the function g to a finite set of S elements $\{x_1, \dots, x_k, \dots, x_S\}$ leads to a matrix of dissimilarities (or semi-distances, or distances) between pairs of the elements. This matrix is said to be Euclidean if and only if one can find S points M_k ($k = 1, \dots, S$) that can be embedded in a Euclidean space so that the Euclidean distance between M_k and M_l is

$$\begin{aligned} g(x_k, x_l) &= \|M_k M_l\| \\ &= \sqrt{(c_k - c_l)^t (c_k - c_l)}, \end{aligned}$$

where c_k and c_l are the vectors of coordinates for M_k and M_l , respectively, in the Euclidean space. These vectors of coordinates can be obtained by a principal coordinate analysis. Consequently, the interest of this Euclidean property is the direct association between the dissimilarities and the obtention of a typology, a graphical representation of the dissimilarities among elements. Other types of graphical displays can be obtained with any dissimilarity coefficient by hierarchical cluster analysis and nonmetric multidimensional scaling.

Examples

Example 1

Let E be the Euclidean space $\mathbb{R}^n$, vector space of all n -tuples of real numbers $(x_1, \dots, x_i, \dots, x_n)$. An element of this space is noted x_k . In that case, each element may be characterized by n quantitative variables $X_1, \dots, X_i, \dots, X_n$. Let $\mathbf{x}_k = (x_{1k}, \dots, x_{ik}, \dots, x_{nk})^t$ and $\mathbf{x}_l = (x_{1l}, \dots, x_{il}, \dots, x_{nl})^t$ be two vectors containing the values taken by the objects k and l , respectively,

for each of the variables considered; $x_k, x_l \in \mathbb{R}^n$. The following dissimilarity coefficients can be used to measure the difference between the objects k and l :

- the Euclidean metric

$$g_1(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t (\mathbf{x}_k - \mathbf{x}_l)};$$

- the Jöreskog distance

$$g_2(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t V^{-1} (\mathbf{x}_k - \mathbf{x}_l)},$$

where $V = \text{diag}(V(Y_1), \dots, V(Y_i), \dots, V(Y_n))$ is the diagonal matrix containing the variances of the n variables

- the Mahalanobis distance

$$g_3(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t W^{-1} (\mathbf{x}_k - \mathbf{x}_l)},$$

where W is the variance-covariance matrix for the n variables.

All of these dissimilarity coefficients are metrics and provide Euclidean dissimilarity matrices.

Example 2

Let E be the set of frequency vectors

$$\begin{aligned} E = \left\{ \mathbf{p} = (p_1, \dots, p_k, \dots, p_S) \mid p_k \geq 0, \right. \\ \left. \sum_{k=1}^S p_k = 1 \right\}. \end{aligned}$$

In that case, let $\mathbf{p}$ and $\mathbf{q}$ be two vectors from E .

Several functions can be used to measure the dissimilarity between the two frequency vectors:

- the Euclidean metric $g_1(\mathbf{p}, \mathbf{q})$
- the taxicab metric, also called Manhattan distance

$$g_4(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^S |p_k - q_k|$$

Modifications of these functions have been proposed in genetics and ecology so that their values lie between 0 and 1:

- the Rogers distance

$$g_5(\mathbf{p}, \mathbf{q}) = \sqrt{\frac{1}{2}(\mathbf{p} - \mathbf{q})'(\mathbf{p} - \mathbf{q})}$$

- the minimal distance from Nei

$$g_6(\mathbf{p}, \mathbf{q}) = \frac{1}{2}(\mathbf{p} - \mathbf{q})'(\mathbf{p} - \mathbf{q})$$

- the absolute genetic distance from Gregorius

$$g_7(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{k=1}^S |p_k - q_k|.$$

The dissimilarity coefficients g_4 , g_5 , and g_7 are metrics, but g_6 is not because it does not verify the triangle inequality (Property C4).

Other dissimilarity coefficients have been developed exclusively for frequency vectors. In 1946, Bhattacharyya introduced the notion of angular distance, considering two multinomial sets characterized by two frequency vectors $\mathbf{p}$ and $\mathbf{q}$. The two vectors $(\sqrt{p_1}, \dots, \sqrt{p_s})$ and $(\sqrt{q_1}, \dots, \sqrt{q_s})$ can be considered as the directions of two lines starting from the origin of a multidimensional space and separated by an angle θ whose cosine is

$$\cos \theta = \sum_{k=1}^S \sqrt{p_k q_k}.$$

The coefficient of dissimilarity proposed by Bhattacharyya is the squared value of this angle:

$$g_8(\mathbf{p}, \mathbf{q}) = \theta^2 = \left[\cos^{-1} \left(\sum_{k=1}^S \sqrt{p_k q_k} \right) \right]^2.$$

Another series of dissimilarity coefficients stems from the probability of drawing two similar objects

from two populations with respective frequency vectors $\mathbf{p}$ and $\mathbf{q}$:

$$\sum_{k=1}^S p_k q_k.$$

Among the dissimilarity coefficients developed from this probability are Nei dissimilarity index in genetics

$$g_9(\mathbf{p}, \mathbf{q}) = -\ln \left(\frac{\sum_{k=1}^S p_k q_k}{\sqrt{\sum_{k=1}^S p_k^2} \sqrt{\sum_{k=1}^S q_k^2}} \right),$$

and Manly overlap index in ecology

$$g_{10}(\mathbf{p}, \mathbf{q}) = 1 - \frac{\sum_{k=1}^S p_k q_k}{\sqrt{\sum_{k=1}^S p_k^2} \sqrt{\sum_{k=1}^S q_k^2}}.$$

Example 3

Let E be the set of binary vectors

$$E = \{\mathbf{u} = (u_1, \dots, u_k, \dots, u_s) \mid u_k \in \{0, 1\}\}.$$

In that case, many dissimilarity coefficients, whose values lie between 0 and 1, have been developed in ecology where a vector $\mathbf{u}$ gives the presence/absence of species in a community. They have been defined from similarity coefficients. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors from E . For each position in the vectors—that is to say, for $k = 1, \dots, S$ —the coefficients look at the similarity of the values taken by $\mathbf{u}$ and $\mathbf{v}$: u_k and v_k . Let a be the number of positions, for $k = 1, \dots, S$, where $u_k = 1$ and $v_k = 1$, b the number of positions where $u_k = 1$ and $v_k = 0$, c the number of positions where $u_k = 0$ and $v_k = 1$, and d the number of positions where $u_k = 0$ and $v_k = 0$. The most used similarity coefficient is Jaccard similarity coefficient

$$s_{11}(\mathbf{u}, \mathbf{v}) = \frac{a}{a + b + c}.$$

A modification of s_{11} , the Sokal and Michener similarity coefficient

$$s_{18}(\mathbf{u}, \mathbf{v}) = \frac{a + d}{a + b + c + d},$$

takes into account the number of species absent from the two communities compared but present in other, comparable communities.

Two modifications of coefficients s_{11} and s_{12} , the Sokal and Sneath similarity coefficient

$$s_{13}(\mathbf{u}, \mathbf{v}) = \frac{a}{a + 2(b + c)}$$

and the Rogers and Tanimoto similarity coefficient

$$s_{14}(\mathbf{u}, \mathbf{v}) = \frac{a + d}{a + 2(b + c) + d},$$

give to the difference (measures b and c) twice as much weight as to the similitude (measures a and d) between the two communities compared. The most common dissimilarity coefficient associated to each similarity coefficient is equal to $g = 1 - s$. It is metric for $g_{11} = 1 - s_{11}$, $g_{12} = 1 - s_{12}$, $g_{13} = 1 - s_{13}$, and $g_{14} = 1 - s_{14}$. For all of these coefficients, a matrix of dissimilarities calculated by $g^* = \sqrt{1 - s}$ is Euclidean.

Dissimilarity and Diversity Coefficients: Rao's Unified Approach

The concepts of dissimilarity and diversity have been linked together by C. R. Rao in a unified theoretical framework. The diversity is the character of objects that exhibit variety. A population in which the objects are numerous and different possesses variety. This variety depends on the relative abundance of the objects and the dissimilarity among these objects. This assertion is at the root of Rao's unified approach.

Consider a population i of S elements $\{x_1, \dots, x_k, \dots, x_S\}$ from any set E . Suppose that these elements are distributed in the population i according to the frequency vector $\mathbf{p}_i = (p_{1i}, \dots, p_{ki}, \dots, p_{Si})^t$. One can calculate $\Delta = [\delta_{kl}]$, $1 \leq k \leq S$, $1 \leq l \leq S$, a

matrix of dissimilarities among the elements, by g , a chosen dissimilarity coefficient: $\delta_{kl} = g(x_k, x_l)$. The diversity in the population i depends on the frequency vector $\mathbf{p}_i$ and the dissimilarity matrix Δ . Rao defined a *diversity coefficient*, also called *quadratic entropy*, as

$$H(\mathbf{p}_i) = \sum_{k=1}^S \sum_{l=1}^S p_{ki} p_{li} \frac{[g(x_k, x_l)]^2}{2}.$$

Consider the matrix $\mathbf{D} = [\delta_{kl}^2/2]$, that is, $\mathbf{D} = [g(x_k, x_l)^2/2]$. Changing for $\mathbf{D}$ in the notations leads to

$$H(\mathbf{p}_i) = \mathbf{p}_i^t \mathbf{D} \mathbf{p}_i.$$

This coefficient has the special feature of being associated with the Jensen difference, a dissimilarity coefficient that calculates dissimilarities among two populations:

$$\begin{aligned} J(\mathbf{p}_i, \mathbf{p}_j) &= 2H\left(\frac{\mathbf{p}_i + \mathbf{p}_j}{2}\right) - H(\mathbf{p}_i) - H(\mathbf{p}_j) \\ &= -\frac{1}{2}(\mathbf{p}_i - \mathbf{p}_j)^t \mathbf{D} (\mathbf{p}_i - \mathbf{p}_j). \end{aligned}$$

The sign of J depends on g via $\mathbf{D}$. It is positive if g is a metric leading to Euclidean matrices. In addition, where g is a metric leading to Euclidean matrices, the dissimilarity coefficient is

$$f(\mathbf{p}_i, \mathbf{p}_j) = \sqrt{2J(\mathbf{p}_i, \mathbf{p}_j)}.$$

Interestingly, f , as with g , is in that case a metric leading to Euclidean matrices. Consequently, the coefficients g and f are measures of inertia (dispersion of points) in a Euclidean space. This result is the heart of the new ordination method called double principal coordinate analysis. It allows a graphical representation of the dissimilarities among populations (coefficient f), together with a projection of the constituting elements.

Thus, the coefficients g and f are connected in a framework of diversity decomposition. The total diversity over several populations is equal to the sum of the average diversity within populations and the diversity among populations. Each component of the decomposition is measured by the quadratic entropy,

but its formula depends either on g when it represents diversity among elements (total and intradiversity) or on f when it represents diversity among populations (interdiversity).

Consider r populations. A weight μ_i is attributed to population i so that $\sum_{i=1}^r \mu_i = 1$. The diversity and dissimilarity coefficients are connected in the following diversity decomposition:

$$H\left(\sum_{i=1}^r \mu_i \mathbf{p}_i\right) = \sum_{i=1}^r \mu_i H(\mathbf{p}_i) + \sum_{i=1}^r \sum_{j=1}^r \mu_i \mu_j \frac{f(\mathbf{p}_i, \mathbf{p}_j)^2}{2}.$$

The component

$$H\left(\sum_{i=1}^r \mu_i \mathbf{p}_i\right) = \sum_{k=1}^S \sum_{l=1}^S \left(\sum_{i=1}^r \mu_i p_{ki}\right) \left(\sum_{i=1}^r \mu_i p_{li}\right) \frac{[g(x_k, x_l)]^2}{2}$$

stems from the total diversity irrespective of populations. It is measured by the quadratic entropy from g and the global frequencies of the S objects. The mean

$$\sum_{i=1}^r \mu_i H(\mathbf{p}_i) = \sum_{i=1}^r \mu_i \left[\sum_{k=1}^S \sum_{l=1}^S p_{ki} p_{li} \frac{g(x_k, x_l)^2}{2} \right]$$

is the average diversity within populations, also measured by the quadratic entropy from the dissimilarity coefficient g and from the frequencies of the objects within populations. Finally, the last term

$$\sum_{i=1}^r \sum_{j=1}^r \mu_i \mu_j \frac{f(\mathbf{p}_i, \mathbf{p}_j)^2}{2}$$

denotes the diversity among populations and is measured by the quadratic entropy from the dissimilarity coefficient f and the relative weights attributed to populations.

This general framework has two interesting specific cases.

1. Where E is the set of values taken by a qualitative variable X , and

$$\frac{g(x_k, x_l)^2}{2} = \begin{cases} 1 & \text{if } x_k \neq x_l \\ 0 & \text{if } x_k = x_l \end{cases},$$

then

$$H(\mathbf{p}_i) = 1 - \sum_{k=1}^S p_{ki}^2,$$

which is known as the Gini-Simpson diversity index, and

$$f(\mathbf{p}_i, \mathbf{p}_j) = \sqrt{(\mathbf{p}_i - \mathbf{p}_j)^t (\mathbf{p}_i - \mathbf{p}_j)}$$

is the Euclidean distance between $\mathbf{p}_i$ and $\mathbf{p}_j$.

2. Where E is the set of values taken by a quantitative variable X , and

$$g(x_k, x_l) = |x_k - x_l|, \quad (1)$$

then

$$H(\mathbf{p}_i) = \sum_{k=1}^S p_{ki} \left(x_k - \sum_{k=1}^S p_{ki} x_k \right)^2,$$

which is the variance of the quantitative variable X . Let $\mathbf{x}$ be the vector $(x_1, \dots, x_k, \dots, x_S)^t$,

$$f(\mathbf{p}_i, \mathbf{p}_j) = \sqrt{(\mathbf{p}_i' \mathbf{x} - \mathbf{p}_j' \mathbf{x})^t (\mathbf{p}_i' \mathbf{x} - \mathbf{p}_j' \mathbf{x})},$$

which can be written simply as the absolute difference between two means

$$f(\mathbf{p}_i, \mathbf{p}_j) = \left| \sum_{k=1}^S p_{ki} x_k - \sum_{k=1}^S p_{kj} x_k \right|. \quad (2)$$

This second writing highlights the consistency between coefficient g (Equation 1), measuring distances between elements, and coefficient f (Equation 2), measuring distances between populations.

In conclusion, the dissimilarity coefficients are functions that may correspond to inertia in Euclidean spaces provided that they verify additional properties. They are used in many disciplines and fit in perfectly with any diversity studies.

—Sandrine Pavoine

See also Distance

Further Reading

- Chessel, D., Dufour, A.-B., & Thioulouse, J. (2004). The ade4 package-I: One-table methods. *R News*, 4, 5–10.
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.
- Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd English ed.). Amsterdam: Elsevier Science.
- Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.
- Pavoine, S., Dufour, A. B., & Chessel, D. (2004). From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis. *Journal of Theoretical Biology*, 228, 523–537.
- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21, 24–43.
- ade4 package for R: <http://pbil.univ-lyon1.fr/R/rplus/ade4dsR.html> (enables you to enter data and compute dissimilarity coefficients, diversity coefficients, the Principal Coordinates Analysis and the double Principal Coordinates Analysis)

DISTANCE

The notion of distance is essential because many statistical techniques are equivalent to the analysis of a specific distance table. For example, principal component analysis and metric multidimensional scaling analyze Euclidean distances, correspondence analysis deals with a χ^2 distance matrix, and discriminant analysis is equivalent to using a Mahalanobis distance. To define a distance is equivalent to defining

rules to assign positive numbers between *pairs* of objects. The most important distances for statistics are Euclidean, generalized Euclidean (which include χ^2 and Mahalanobis), Minkowsky (which include the sorting and the symmetric difference distances), and the Hellinger distances.

Notation and Definition

For convenience, we restrict our discussion to distance between vectors because they are the objects mostly used in statistics. Let **a**, **b**, and **c** be three vectors with J elements each. A distance is a function that associates to any pair of vectors a real positive number, denoted $d(\mathbf{a}, \mathbf{b})$, which has the following properties:

$$d(\mathbf{a}, \mathbf{a}) = 0 \quad (1)$$

$$d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a}) \text{ [symmetry]} \quad (2)$$

$$d(\mathbf{a}, \mathbf{b}) = d(\mathbf{a}, \mathbf{c}) + d(\mathbf{c}, \mathbf{b}) \text{ [triangular inequality]} \quad (3)$$

A Minimalist Example: The Sorting Distance

The axioms defining a distance are very easily met. For example, suppose that we consider two objects and assign the number 1 if we find them different and 0 if we find them alike. This procedure defines a distance called the *sorting* distance because the number assigned to a pair of same objects will be equal to 0 (this satisfies Axiom 1). Axiom 2 is also satisfied because the order of the objects is irrelevant. For the third axiom, we need to consider two cases, if $d(\mathbf{a}, \mathbf{b})$ is equal to 0, the sum $d(\mathbf{a}, \mathbf{c}) + d(\mathbf{c}, \mathbf{b})$ can take only the values 0, 1, or 2, which will all satisfy Axiom 3. If **a** and **b** are different, $d(\mathbf{a}, \mathbf{b})$ is equal to 1 and **c** cannot be identical to *both* of them, and therefore the sum $d(\mathbf{a}, \mathbf{c}) + d(\mathbf{c}, \mathbf{b})$ can take only the values 1 or 2, which will both satisfy Axiom 3.

With the same argument, we can see that if we ask a set of respondents to sort objects into piles, the number of participants who do not sort two objects together defines a distance between the sorted objects.

The Euclidean Distance

The most well-known distance is the *Euclidean distance*, which is defined as

$$\begin{aligned} d(\mathbf{a}, \mathbf{b}) &= \|\mathbf{a} - \mathbf{b}\| = \sqrt{(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b})} \\ &= \sqrt{\sum_j (a_j - b_j)^2} \end{aligned} \quad (4)$$

(with $\|\mathbf{a}\|$ being the norm of $\mathbf{a}$, and a_j and b_j being the j th element of $\mathbf{a}$ and $\mathbf{b}$). Expressed as a squared distance (in a Euclidean world, it is always more practical to work with squared quantities because of the Pythagorean theorem), it is computed as

$$d^2(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b}). \quad (5)$$

For example, with

$$\mathbf{a} = \begin{bmatrix} 2 \\ 5 \\ 10 \\ 20 \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad (6)$$

the vector $\mathbf{a} - \mathbf{b}$ gives

$$\mathbf{a} - \mathbf{b} = \begin{bmatrix} 2 - 1 \\ 5 - 2 \\ 10 - 3 \\ 20 - 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 7 \\ 16 \end{bmatrix}, \quad (7)$$

and

$$\begin{aligned} d^2(\mathbf{a}, \mathbf{b}) &= (\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b}) \\ &= \sum_{j=1}^4 (a_j - b_j)^2 \\ &= 1^2 + 3^2 + 7^2 + 16^2 \\ &= 315. \end{aligned} \quad (8)$$

The Euclidean distance between two vectors can also be expressed via the notion of scalar product and cosine between vectors. By developing Equation 5 for the distance between vectors, we find that

$$\begin{aligned} d^2(\mathbf{a}, \mathbf{b}) &= (\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b}) \\ &= \mathbf{a}^T \mathbf{a} + \mathbf{b}^T \mathbf{b} - 2\mathbf{a}^T \mathbf{b} \\ &= \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\| \times \|\mathbf{b}\| \times \cos(\mathbf{a}, \mathbf{b}). \end{aligned} \quad (9)$$

In the particular case of vectors with a unit norm, the distance between $\mathbf{a}$ and $\mathbf{b}$ simplifies to

$$d^2(\mathbf{a}, \mathbf{b}) = 2[1 - \cos(\mathbf{a}, \mathbf{b})]. \quad (10)$$

When two vectors are centered (i.e., when their mean is equal to 0), their cosine is equal to the coefficient of correlation. This shows that we can define a (Euclidean) distance between two series of numbers as 1 minus their correlation.

Generalized Euclidean

The Euclidean distance can be generalized by taking into account constraints expressed by a matrix conformable with the vectors. Specifically, let $\mathbf{W}$ denote a $J \times J$ positive definite matrix, the generalized Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$ becomes

$$d_{\mathbf{W}}^2(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})^T \mathbf{W} (\mathbf{a} - \mathbf{b}). \quad (11)$$

The most well-known generalized Euclidean distances are the χ^2 and the Mahalanobis distances.

χ^2 Distance

The χ^2 distance is associated with correspondence analysis. It is a distance between profiles. Recall that a vector is called a *profile* when it is composed of numbers greater than or equal to zero whose sum is equal to 1 (such a vector is sometimes called a *stochastic* vector). The χ^2 distance is defined for the rows (or the columns after transposition of the data table) of a contingency table such as the one shown in Table 1. The first step of the computation of the distance is to transform the rows into row profiles, which is done by dividing each row by its total. There are I rows and J columns in a contingency table. The *mass* of each row is denoted r_i , and the mass vector $\mathbf{r}$. The barycenter of the rows, denoted $\mathbf{c}$ is computed by transforming the total of the columns into a row profile. It can also be computed as the weighted average of the row profiles (with the weights being given by the mass vector $\mathbf{r}$). For the χ^2 distance, the $\mathbf{W}$ matrix is diagonal, which is equivalent to assigning a weight to each column.

This weight is equal to the inverse of the relative frequency of the column. This is expressed formally by expressing $\mathbf{W}$ as

$$\mathbf{W} = (\text{diag}\{\mathbf{c}\})^{-1}. \quad (12)$$

With this coding schema, variables that are used often contribute less to the distance between rows than variables that are used rarely. For example, from Table 1, we find that the weight matrix is equal to

$$\begin{aligned} \mathbf{W} &= \mathbf{D}_w = \text{diag}\{\mathbf{w}\} \\ &= \begin{bmatrix} .2973^{-1} & 0 & 0 \\ 0 & .5642^{-1} & 0 \\ 0 & 0 & .1385^{-1} \end{bmatrix} \\ &= \begin{bmatrix} 3.3641 & 0 & 0 \\ 0 & 1.7724 & 0 \\ 0 & 0 & 7.2190 \end{bmatrix}. \end{aligned} \quad (13)$$

For example, the χ^2 distance between Rousseau and Chateaubriand is equal to

$$\begin{aligned} d^2(\text{Rousseau}, \text{Chateaubriand}) &= 3.364 \times (.291 - .270)^2 + 1.772 \\ &\times (.486 - .526)^2 + 7.219 \times (.223 - .214)^2 \\ &= .0036. \end{aligned} \quad (14)$$

This distance is called the χ^2 distance because the sum of the weighted distances from the rows to their barycenter is proportional to the χ^2 computed to test the independence of the rows and the columns of the table. Formally, if we denoted by N the grand total of the contingency table, by $d^2(i, g)$ the distance from row i to the barycenter of the table, and by $d^2(i, i')$ the distance from row i to row i' , we obtain the following equality:

$$\sum_i r_i d^2(i, g) = \sum_{i>i'} r_i r_{i'} d^2(i, i') = \frac{1}{N} \chi^2. \quad (15)$$

The metric multidimensional scaling analysis of a χ^2 distance matrix (with masses given by $\mathbf{r}$) is equivalent to correspondence analysis.

Mahalanobis Distance

The Mahalanobis distance is defined between rows of a table. The weight matrix is obtained as the inverse of the columns variance/covariance matrix. Formally, if we denoted by $\mathbf{S}$ the variance/covariance matrix between the columns of a data table, the weight matrix of the Mahalanobis distance is defined as $\mathbf{W} = \mathbf{S}^{-1}$.

Table 1 Data for the Computation of the χ^2 , Mahalanobis, and Hellinger Distances

Author's name	Raw Data			$N \times \mathbf{r}$	$\mathbf{r}$	Row Profiles		
	Period	Comma	Other			Period	Comma	Other
Rousseau	7836	13112	6026	26974	.0189	.2905	.4861	.2234
Chateaubriand	53655	102383	42413	198451	.1393	.2704	.5159	.2137
Hugo	115615	184541	59226	359382	.2522	.3217	.5135	.1648
Zola	161926	340479	62754	565159	.3966	.2865	.6024	.1110
Proust	38177	105101	12670	155948	.1094	.2448	.6739	.0812
Giraudoux	46371	58367	14299	119037	.0835	.3896	.4903	.1201
$\sum Nc^T$	423580	803983	197388	1424951	1.0000			
c^T	.2973	.5642	.1385					
w^T	3.3641	1.7724	7.2190					

Notes: The punctuation marks of six French writers (from Abdi & Valentin, 2006). The column labeled $N \times \mathbf{r}$ gives the total number of punctuation marks used by each author. The mass of each row is the proportion of punctuation marks used by this author. The row labeled $N \times c^T$ gives the total of each column. This is the total number of times this punctuation mark was used. The centroid row (or barycenter, or center of gravity), gives the proportion of each punctuation mark in the sample. The weight of each column is the inverse of the centroid.

Using, again, the data from Table 1, we obtain

$$\mathbf{S} = 10^{10} \times \begin{bmatrix} 0.325 & 0.641 & 0.131 \\ 0.641 & 1.347 & 0.249 \\ 0.131 & 0.249 & 0.063 \end{bmatrix} \text{ and } \mathbf{S}^{-1} = 10^{-7} \times \begin{bmatrix} 0.927 & -0.314 & -0.683 \\ -0.314 & 0.134 & 0.124 \\ -0.683 & 0.124 & 1.087 \end{bmatrix}. \quad (16)$$

With these values, we find that the Mahalanobis distance between Rousseau and Chateaubriand is equal to

$$d^2(\text{Rousseau, Chateaubriand}) = 10^{-7} \times \left(\begin{bmatrix} -45819 \\ -89271 \\ -36387 \end{bmatrix}^T \times \begin{bmatrix} 0.927 & -0.314 & -0.683 \\ -0.314 & 0.134 & 0.124 \\ -0.683 & 0.124 & 1.087 \end{bmatrix} \times \begin{bmatrix} -45819 \\ -89271 \\ -36387 \end{bmatrix} \right) \quad (17)$$

$$\approx 4.0878.$$

The Mahalanobis distance can be seen as a multivariate equivalent of the z-score transformation. The metric multidimensional scaling analysis of a Mahalanobis distance matrix is equivalent to discriminant analysis.

Minkowski's Distance

The Euclidean distance is a particular case of the more general family of Minkowski's distances. The p distance (or a Minkowski's distance of degree p), between two vectors is defined as

$$\begin{aligned} \mathbf{a} &= [a_1, \dots, a_j, \dots, a_J]^T \text{ and} \\ \mathbf{b} &= [b_1, \dots, b_j, \dots, b_J]^T \end{aligned} \quad (18)$$

as

$$d_p(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_p = \left[\sum_j |a_j - b_j|^p \right]^{\frac{1}{p}}. \quad (19)$$

The most frequently used Minkowski's distances are the distances of degree 1, 2, and ∞ . A distance of degree 1 is also called the *city-block* or *taxicab* distance. When the vectors are binary numbers (i.e., 1 and 0), the elements of the vector code for membership to a set (i.e., 1 means the element belongs to the set, 0 means it does not). In this case, the degree 1 distance is commonly referred to as the *Hamming distance* or the *symmetric difference distance*. (The symmetric difference distance is a set operation that associates to two sets a new set made of the elements

of these sets that belong to only one of them—i.e., elements that belong to both sets are excluded. The symmetric difference distance gives the number of the elements of the symmetric difference set.)

When p is equal to 2, we obtain the usual Euclidean distance. With $p = \infty$, we take the largest absolute value of the difference between the vectors as defining the distance between vectors.

For example, with the vectors

$$\mathbf{a} = \begin{bmatrix} 2 \\ 5 \\ 10 \\ 20 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad (20)$$

the Minkowski distance of degree 1 is

$$d_1(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^4 |a_j - b_j| = 1 + 3 + 7 + 16 = 27, \quad (21)$$

and the Minkowski distance of degree ∞ is

$$d_\infty(\mathbf{a}, \mathbf{b}) = \max_j |a_j - b_j| = \max\{1, 3, 7, 16\} = 16. \quad (22)$$

Hellinger

The Hellinger distance is defined between vectors having only positive or zero elements. In general (like the χ^2 distance), it is used for row profiles. The Hellinger distance between vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$d(\mathbf{a}, \mathbf{b}) = \left[\sum_j^J \left(\sqrt{a_j} - \sqrt{b_j} \right)^2 \right]^{\frac{1}{2}}. \quad (23)$$

Because the Hellinger distance is not sensitive to discrepancies between columns, it is sometimes used as an alternative to the χ^2 distance. An interesting property of the Hellinger distance when applied to row profiles is that the vectors representing these profiles can be represented as points on a sphere (or hypersphere when the number of elements of the vector is larger than 3).

For our example, we find that the Hellinger distance between the row profiles of Rousseau and Chateaubriand is equal to

$$\begin{aligned} d^2(\text{Rousseau}, \text{Chateaubriand}) &= \left[\left(\sqrt{2905} - \sqrt{.2704} \right)^2 \right. \\ &\quad + \left(\sqrt{.4861} - \sqrt{.5259} \right)^2 \\ &\quad \left. + \left(\sqrt{.2234} - \sqrt{.2137} \right)^2 \right]^{\frac{1}{2}} \\ &= .0302. \end{aligned} \quad (24)$$

How to Analyze Distance Matrices

Distance matrices are often computed as the first step of data analysis. In general, distance matrices are analyzed by finding a convenient graphic representation for their elements. These representations approximate the original distance by another distance such as (a) a low-dimensionality Euclidean distance (e.g., multidimensional scaling, DISTATIS); or (b) a graph (e.g., cluster analysis, additive tree representations).

—Hervé Abdi

See also Correspondence Analysis; Discriminant Analysis; Discriminant Correspondence Analysis; Dissimilarity Coefficient; DISTATIS; Metric Multidimensional Scaling

Further Reading

- Abdi, H. (1990). Additive-tree representations. *Lecture Notes in Biomathematics*, 84, 43–59.
- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H., & Valentin, D. (2006). *Mathématiques pour les sciences cognitives* (Mathematics for cognitive sciences). Grenoble, France: Presses Universitaires de Grenoble.
- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistiques Appliquées*, 26, 29–37.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Rao, C. R. (1995). Use of Hellinger distance in graphical displays. In E.-M. Tiit, T. Kollo, & H. Niemi (Eds.), *Multivariate statistics and matrices in statistics* (pp. 143–161). Leiden, Netherlands: Brill.

DISTATIS

DISTATIS is a generalization of classical multidimensional scaling (MDS) proposed by Abdi, Valentin, O'Toole, and Edelman. Its goal is to analyze several distance matrices computed on the same set of objects. The name DISTATIS is derived from a technique called STATIS, whose goal is to analyze multiple data sets. DISTATIS first evaluates the similarity between distance matrices. From this analysis, a compromise matrix is computed that represents the best aggregate of the original matrices. The original distance matrices are then projected onto the compromise.

The data sets to analyze are distance matrices obtained on the same set of objects. These distance matrices may correspond to measurements taken at different times. In this case, the first matrix corresponds to the distances collected at time $t = 1$, the second one to the distances collected at time $t = 2$, and so on. The goal of the analysis is to evaluate if the relative positions of the objects are stable over time. The different matrices, however, do not need to represent time. For example, the distance matrices can be derived from different methods. The goal of the analysis, then, is to evaluate if there is an agreement between the methods.

The general idea behind DISTATIS is first to transform each distance matrix into a cross-product matrix as it is done for a standard MDS. Then, these cross-product matrices are aggregated to create a compromise cross-product matrix that represents their consensus. The compromise matrix is obtained as a weighted average of individual cross-product matrices. The principal component analysis (PCA) of the compromise gives the position of the objects in the compromise space. The position of the object for each study can be represented in the compromise space as supplementary points. Finally, as a by-product of the weight computation, the studies can be represented as points in a multidimensional space.

An Example

To illustrate DISTATIS, we will use the set of faces displayed in Figure 1. Four different “systems” or algorithms are compared, each of them computing a distance matrix between the faces. The first system corresponds to PCA and computes the squared Euclidean distance between faces directly from the pixel values of the images. The second system starts by taking measurements on the faces (see Figure 2) and computes the squared Euclidean distance between faces from these measures. The third distance matrix is obtained by first asking human observers to rate the faces on several dimensions (e.g., beauty, honesty,

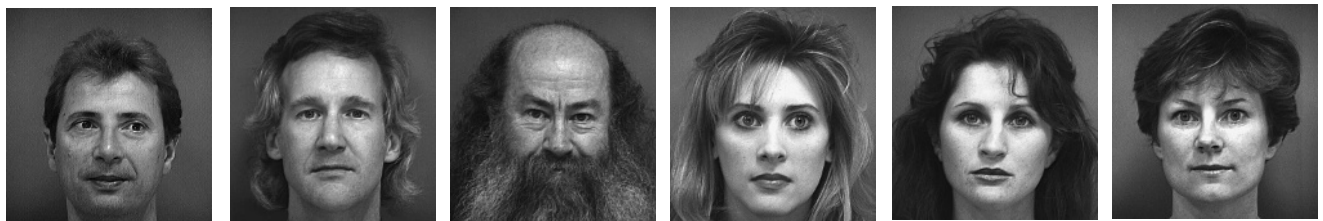


Figure 1 Six Faces to Be Analyzed by Different “Algorithms”

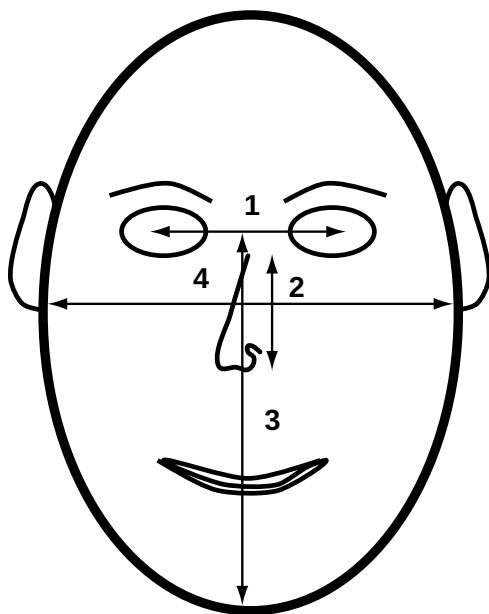


Figure 2 The Measures Taken on a Face

empathy, and intelligence) and then computing the squared Euclidean distance from these measures. The fourth distance matrix is obtained from pairwise similarity ratings (on a scale from 1 to 7) collected from human observers. The average similarity rating s was transformed into a distance using Shepard’s transformation: $d = \exp\{-s^2\}$.

Notations

The raw data consist of T data sets and we will refer to each data set as a *study*. Each study is an $I \times I$ distance matrix denoted $\mathbf{D}_{[t]}$, where I is the number of objects and t denotes the study.

Here, we have $T = 4$ studies. Each study corresponds to a 6×6 distance matrix as shown below.

Study 1 (Pixels):

$$\mathbf{D}_{[1]} = \begin{bmatrix} 0 & .112 & .148 & .083 & .186 & .110 \\ .112 & 0 & .152 & .098 & .158 & .134 \\ .146 & .152 & 0 & .202 & .285 & .249 \\ .083 & .098 & .202 & 0 & .131 & .110 \\ .186 & .158 & .285 & .131 & 0 & .155 \\ .110 & .134 & .249 & .110 & .155 & 0 \end{bmatrix}.$$

Study 2 (Measures):

$$\mathbf{D}_{[2]} = \begin{bmatrix} 0 & 0.60 & 1.98 & 0.42 & 0.14 & 0.58 \\ 0.60 & 0 & 2.10 & 0.78 & 0.42 & 1.34 \\ 1.98 & 2.10 & 0 & 2.02 & 1.72 & 2.06 \\ 0.42 & 0.78 & 2.02 & 0 & 0.50 & 0.88 \\ 0.14 & 0.42 & 1.72 & 0.50 & 0 & 0.30 \\ 0.58 & 1.34 & 2.06 & 0.88 & 0.30 & 0 \end{bmatrix}.$$

Study 3 (Ratings):

$$\mathbf{D}_{[3]} = \begin{bmatrix} 0 & 0.54 & 1.39 & 5.78 & 10.28 & 6.77 \\ 0.54 & 0 & 1.06 & 3.80 & 6.83 & 4.71 \\ 1.39 & 1.06 & 0 & 8.01 & 11.03 & 5.72 \\ 5.78 & 3.80 & 8.01 & 0 & 2.58 & 6.09 \\ 10.28 & 6.83 & 11.03 & 2.58 & 0 & 3.53 \\ 6.77 & 4.71 & 5.72 & 6.09 & 3.53 & 0 \end{bmatrix}.$$

Study 4 (Pairwise):

$$\mathbf{D}_{[4]} = \begin{bmatrix} 0 & .014 & .159 & .004 & .001 & .002 \\ .014 & 0 & .018 & .053 & .024 & .004 \\ .159 & .018 & 0 & .271 & .067 & .053 \\ .004 & .053 & .271 & 0 & .001 & .008 \\ .001 & .024 & .067 & .001 & 0 & .007 \\ .002 & .004 & .053 & .008 & .007 & 0 \end{bmatrix}.$$

Distance matrices cannot be analyzed directly and need to be transformed. This step corresponds to MDS and transforms a distance matrix into a cross-product matrix.

We start with an $I \times I$ distance matrix $\mathbf{D}$, with an $I \times 1$ vector of mass (whose elements are all positive

or zero and whose sum is equal to 1) denoted $\mathbf{m}$, such that

$$\mathbf{m}^T \mathbf{1} = 1. \quad (1)$$

If all observations have the same mass (as in here) $m_i = \frac{1}{I}$. We then define the centering matrix, which is equal to

$$\mathbf{\Xi} = \mathbf{I} - \frac{1}{I} \mathbf{1} \mathbf{1}^T, \quad (2)$$

and the cross-product matrix denoted by $\tilde{\mathbf{S}}$ is obtained as

$$\tilde{\mathbf{S}} = -\frac{1}{2} \mathbf{\Xi} \mathbf{D} \mathbf{\Xi}^T. \quad (3)$$

For example, the first distance matrix is transformed into the following cross-product matrix:

$$\begin{aligned} \tilde{\mathbf{S}}_{[1]} &= -\frac{1}{2} \mathbf{\Xi} \mathbf{D}_{[1]} \mathbf{\Xi}^T \\ &= \begin{bmatrix} 0.042 & -0.013 & 0.002 & -0.001 & -0.028 & -0.003 \\ -0.013 & 0.045 & 0.000 & -0.007 & -0.012 & -0.013 \\ 0.002 & 0.000 & 0.108 & -0.027 & -0.044 & -0.039 \\ -0.001 & -0.007 & -0.027 & 0.040 & -0.001 & -0.004 \\ -0.028 & -0.012 & -0.044 & -0.001 & 0.088 & -0.002 \\ -0.003 & -0.013 & -0.039 & -0.004 & -0.002 & 0.062 \end{bmatrix}. \end{aligned}$$

In order to compare the studies, we need to normalize the cross-product matrices. There are several possible normalizations; here we normalize the cross-product matrices by dividing each matrix by its first eigenvalue (an idea akin to multiple factor analysis). The first eigenvalue of matrix $\tilde{\mathbf{S}}_{[1]}$ is equal to $\lambda_1 = .16$, and matrix $\tilde{\mathbf{S}}_{[1]}$ is transformed into a normalized cross-product matrix denoted $\mathbf{S}_{[1]}$ as

$$\begin{aligned} \mathbf{S}_{[1]} &= \lambda_1^{-1} \times \tilde{\mathbf{S}}_{[1]} \\ &= \begin{bmatrix} .261 & -.079 & .013 & -.003 & -.174 & -.018 \\ -.079 & .280 & .002 & -.042 & -.077 & -.084 \\ .013 & .002 & .675 & -.168 & -.276 & -.246 \\ -.003 & -.042 & -.168 & .249 & -.009 & -.026 \\ -.174 & -.077 & -.276 & -.009 & .552 & -.015 \\ -.017 & -.084 & -.246 & -.026 & -.015 & .388 \end{bmatrix}. \quad (4) \end{aligned}$$

Computing the Compromise Matrix

The *compromise matrix* is a cross-product matrix that gives the best compromise of the studies. It is obtained as a weighted average of the study cross-product matrices. The weights are chosen so that studies agreeing the most with other studies will have the larger weights. To find these weights, we need to analyze the relationships between the studies.

The *compromise matrix* is a cross-product matrix that gives the best compromise of the cross-product matrices representing each study. It is obtained as a weighted average of these matrices. The first step is to derive an optimal set of weights. The principle to find this set of weights is similar to that described for STATIS and involves the following steps.

Comparing the Studies

To analyze the similarity structure of the studies we start by creating a *between-study cosine matrix* denoted $\mathbf{C}$. This is a $T \times T$ matrix whose generic term $c_{t,t'}$ gives the cosine between studies t and t' . This cosine, also known as the R_V coefficient, is defined as

$$R_V = [c_{t,t'}] = \frac{\text{trace} \{ \mathbf{S}_{[t]}^T \mathbf{S}_{[t']} \}}{\sqrt{\text{trace} \{ \mathbf{S}_{[t]}^T \mathbf{S}_{[t]} \} \times \text{trace} \{ \mathbf{S}_{[t']}^T \mathbf{S}_{[t']} \}}}. \quad (5)$$

Using this formula, we get the following matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} 1.00 & .77 & .76 & .40 \\ .77 & 1.00 & .41 & .53 \\ .76 & .41 & 1.00 & .30 \\ .40 & .53 & .30 & 1.00 \end{bmatrix}. \quad (6)$$

PCA of the Cosine Matrix

The cosine matrix has the following eigen-decomposition

$$\mathbf{C} = \mathbf{P}\mathbf{\Theta}\mathbf{P}^T \text{ with } \mathbf{P}^T\mathbf{P} = \mathbf{I}, \quad (7)$$

where $\mathbf{P}$ is the matrix of eigenvectors and $\mathbf{\Theta}$ is the diagonal matrix of the eigenvalues of $\mathbf{C}$. For our example, the eigenvectors and eigenvalues of $\mathbf{C}$ are

$$\mathbf{P} = \begin{bmatrix} .58 & .28 & -.21 & .74 \\ .53 & -.24 & -.64 & -.50 \\ .48 & .56 & .51 & -.44 \\ .40 & -.74 & .53 & .11 \end{bmatrix} \text{ and } \text{diag} \{ \mathbf{\Theta} \} = \begin{bmatrix} 2.62 \\ 0.80 \\ 0.49 \\ 0.09 \end{bmatrix}.$$

An element of a given eigenvector represents the projection of one study on this eigenvector. Thus, the T studies can be represented as points in the eigen-space and their similarities analyzed visually. This step corresponds to a PCA of the between-studies space. In general, when we plot the studies in their factor space, we want to give to each component the length corresponding to its eigenvalue (i.e., the inertia of the coordinates of a dimension is equal to the eigenvalue of this dimension, which is the standard procedure in PCA and MDS). For our example, we obtain the following coordinates:

$$\mathbf{G} = \mathbf{P} \times \mathbf{\Theta}^{\frac{1}{2}} = \begin{bmatrix} .93 & .25 & -.14 & .23 \\ .85 & -.22 & -.45 & -.15 \\ .78 & .50 & .36 & -.13 \\ .65 & -.66 & .37 & .03 \end{bmatrix}.$$

As an illustration, Figure 3 displays the projections of the four algorithms onto the first and second eigenvectors of the cosine matrix.

Because the matrix is not centered, the first eigenvector represents what is common to the different studies. The more similar a study is to the other studies, the more it will contribute to this eigenvector. Or, in other words, studies with larger projections on the first eigenvector are more similar to the other studies than studies with smaller projections. Thus, the elements of the first eigenvector give the optimal weights to compute the compromise matrix.

Computing the Compromise

As for STATIS, the weights are obtained by dividing each element of $\mathbf{p}_1$ by their sum. The vector containing these weights is denoted $\boldsymbol{\alpha}$. For our example, we obtain

$$\boldsymbol{\alpha} = [.29 \ .27 \ .24 \ .20]^T. \quad (8)$$

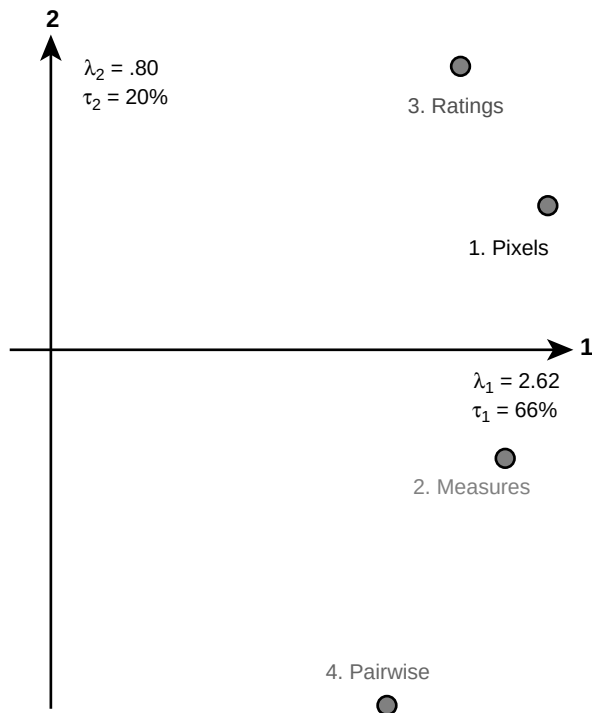


Figure 3 Plot of the Between-Studies Space (i.e., eigenanalysis of the matrix $\mathbf{C}$)

With α_t denoting the weight for the t th study, the compromise matrix, denoted $\mathbf{S}_{[+]}$, is computed as

$$\mathbf{S}_{[+]} = \sum_t \alpha_t \mathbf{S}_{[t]}. \quad (9)$$

In our example, this gives

$$\mathbf{S}_{[+]} = \begin{bmatrix} .176 & .004 & -.058 & .014 & -.100 & -.036 \\ .004 & .178 & .022 & -.038 & -.068 & -.010 \\ -.058 & .022 & .579 & -.243 & -.186 & -.115 \\ .014 & -.038 & -.243 & .240 & .054 & -.027 \\ -.100 & -.068 & -.186 & .054 & .266 & .034 \\ -.036 & -.010 & -.115 & -.027 & .034 & .243 \end{bmatrix}.$$

How Representative Is the Compromise?

To evaluate the quality of the compromise, we need an index of quality. This is given by the first eigenvalue of matrix $\mathbf{C}$, which is denoted ϑ_1 . An alternative index of quality (easier to interpret) is the ratio of the first eigenvalue of $\mathbf{C}$ to the sum of its eigenvalues:

$$\text{Quality of compromise} = \frac{\vartheta_1}{\sum_{\ell} \vartheta_{\ell}} = \frac{\vartheta_1}{\text{trace}\{\mathbf{\Theta}\}}. \quad (10)$$

Here, the quality of the compromise is evaluated as

$$\text{Quality of compromise} = \frac{\vartheta_1}{\text{trace}\{\mathbf{\Theta}\}} = \frac{2.62}{4} \approx .66. \quad (11)$$

So, we can say that the compromise “explains” 66% of the inertia of the original set of data tables. This is a relatively small value, and this indicates that the algorithms differ substantially on the information they capture about the faces.

Analyzing the Compromise

The eigendecomposition of the compromise is

$$\mathbf{S}_{[+]} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \quad (12)$$

with, in our example,

$$\mathbf{Q} = \begin{bmatrix} .017 & .474 & -.451 & -.107 & -.627 \\ .121 & .400 & .256 & .726 & .258 \\ .823 & -.213 & .114 & -.308 & .053 \\ -.388 & .309 & .159 & -.566 & .492 \\ -.348 & -.443 & .549 & .043 & -.462 \\ -.192 & -.527 & -.626 & .211 & .287 \end{bmatrix} \quad (13)$$

and

$$\text{diag}\{\mathbf{\Lambda}\} = [.80 \ .35 \ .26 \ .16 \ .11]^T. \quad (14)$$

From Equations 13 and 14, we can compute the compromise factor scores for the faces as

$$\mathbf{F} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}} = \begin{bmatrix} -.015 & .280 & -.228 & -.043 & -.209 \\ .108 & .236 & .129 & .294 & .086 \\ .738 & -.126 & .058 & -.125 & .018 \\ -.348 & .182 & .080 & -.229 & .164 \\ -.312 & -.262 & .277 & .018 & -.155 \\ -.172 & -.311 & -.316 & .086 & .096 \end{bmatrix}. \quad (15)$$

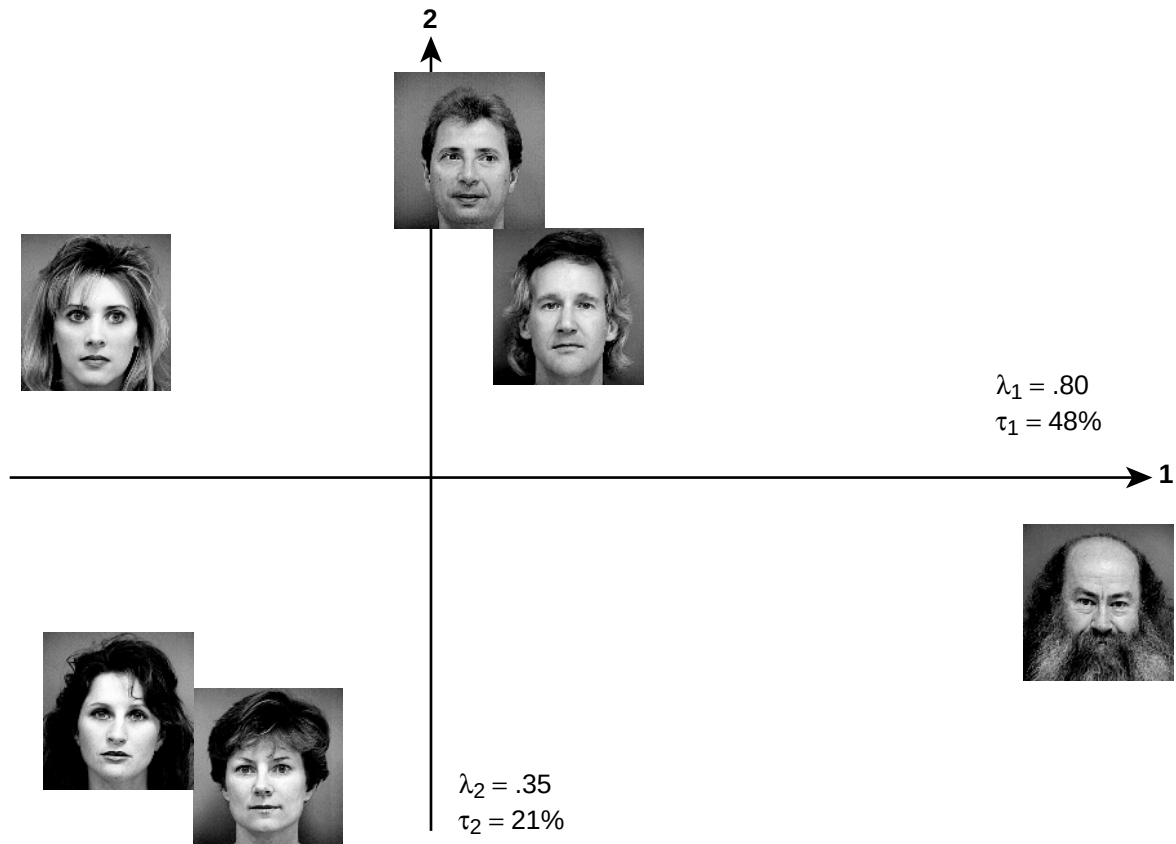


Figure 4 Analysis of the Compromise: Plot of the Faces in the Plane Defined by the First Two Principal Components of the Compromise Matrix

In the $\mathbf{F}$ matrix, each row represents an object (i.e., a face) and each column a component. Figure 4 displays the faces in the space defined by the first two principal components. The first component has an eigenvalue equal to $\lambda_1 = .80$; such a value explains 48% of the inertia. The second component, with an eigenvalue of .35, explains 21% of the inertia. The first component is easily interpreted as the opposition of the male to the female faces (with Face #3 appearing extremely masculine). The second dimension is more difficult to interpret and seems linked to hair color (i.e., light hair vs. dark or no hair).

Projecting the Studies Into the Compromise Space

Each algorithm provided a cross-product matrix, which was used to create the compromise cross-product

matrix. The analysis of the compromise reveals the structure of the face space common to the algorithms. In addition to this common space, we want also to see how each algorithm “interprets” or distorts this space. This can be achieved by projecting the cross-product matrix of each algorithm onto the common space. This operation is performed by computing a projection matrix that transforms the scalar product matrix into loadings. The projection matrix is deduced from the combination of Equations 12 and 15, which gives

$$\mathbf{F} = \mathbf{S}_{[+]} \mathbf{Q} \mathbf{\Lambda}^{-\frac{1}{2}}. \quad (16)$$

This shows that the projection matrix is equal to $(\mathbf{Q} \mathbf{\Lambda}^{-\frac{1}{2}})$. It is used to project the scalar product matrix of each study onto the common space. For example, the coordinates of the projections for the first study are obtained by first computing the matrix

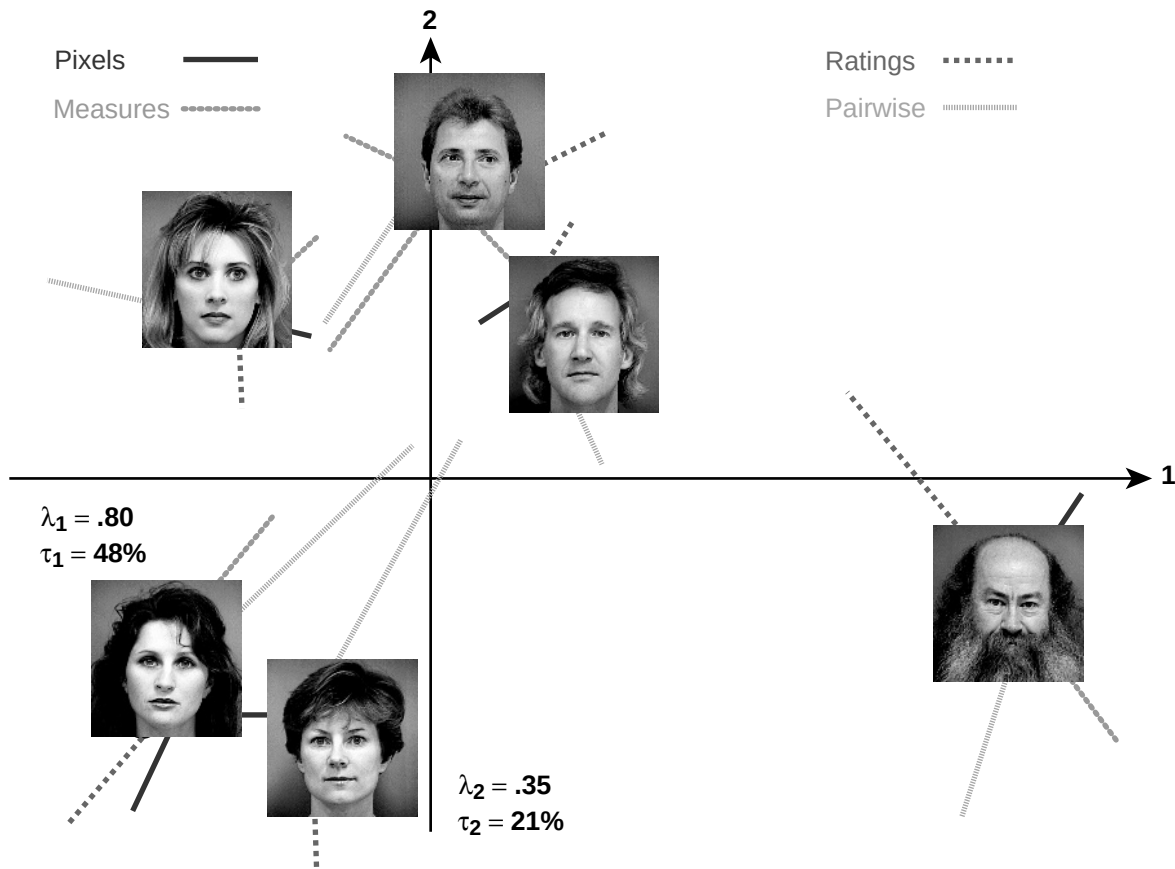


Figure 5 The Compromise: Projection of the Algorithm Matrices Onto the Compromise Space

$$Q\Lambda^{-\frac{1}{2}} = \begin{bmatrix} -0.02 & 0.80 & -0.89 & -0.26 & -1.88 \\ 0.13 & 0.68 & 0.51 & 1.79 & 0.77 \\ 0.92 & -0.36 & 0.23 & -0.76 & 0.16 \\ -0.43 & 0.52 & 0.31 & -1.40 & 1.48 \\ -0.39 & -0.75 & 1.09 & 0.11 & -1.39 \\ -0.21 & -0.89 & -1.24 & 0.52 & 0.86 \end{bmatrix}, \quad (17)$$

and then using this matrix to obtain the coordinates of the projection as

$$\begin{aligned} F_{[1]} &= S_{[1]} \left(Q\Lambda^{-\frac{1}{2}} \right) \\ &= \begin{bmatrix} .07 & .30 & -.44 & -.24 & -.33 \\ .11 & .24 & .22 & .53 & .34 \\ .85 & .11 & .09 & -.44 & .01 \\ -.26 & .19 & .04 & -.31 & .30 \\ -.47 & -.50 & .67 & .18 & -.57 \\ -.30 & -.33 & -.59 & .28 & .25 \end{bmatrix}. \end{aligned} \quad (18)$$

The same procedure is used to compute the matrices of the projections onto the compromise space for the other algorithms.

Figure 5 shows the first two principal components of the compromise space along with the projections of each of the algorithms. The position of a face in the compromise is the barycenter of its positions for the four algorithms. In order to facilitate the interpretation, we have drawn lines linking the position of each face for each of the four algorithms to its compromise position. This picture confirms that the algorithms differ substantially. It shows also that some faces are more sensitive to the differences between algorithms (e.g., compare Faces 3 and 4).

—Hervé Abdi and
Dominique Valentin

See also Distance; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Multiple Factor Analysis, R_v and Congruence Coefficients; STATIS

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H., Valentin, D., O'Toole, A. J., & Edelman, B. (2005). DISTATIS: The analysis of multiple distance matrices. *Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition*, pp. 42–47.
- Escofier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples*. Paris: Dunod.

DIXON TEST FOR OUTLIERS

In statistics, we assume that our data come from some probability model, a hypothetical or ideal model for doing statistical analysis mathematically. In the real-life world, unfortunately, the data in our hands usually have some outliers: outlying observations discordant from the hypothesized model. It is well known that outliers in the data set can severely distort the result of statistical inference.

Generally speaking, there are two strategies to deal with outliers. One way is to use some robust methods to accommodate outliers in the data, such as using the sample median (instead of the sample mean) to estimate a population mean. The other way is to try to identify outliers in the sample and then modify or simply delete them for further data analysis. Our topic here is the identification of outliers in the sample data.

Suppose that our sample data $x_1, x_2, \dots, x_n$ come from an interesting population or distribution. Arrange the sample observations $x_1, x_2, \dots, x_n$ in ascending order: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, which are called order statistics. Suppose that we have k suspicious lower outliers $x_{(1)}, x_{(2)}, \dots, x_{(k)}$ or upper outliers $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$ in the sample (the case of having both lower and upper outliers is too complicated to be considered here), where k (the number of suspicious outliers) is much smaller than the sample size n . We want to test if they are significantly discordant from the rest of sample observations.

Hypothesis Test for Outliers

To test the discordancy of suspicious outliers, we need to propose statistical hypotheses. For example, to test

the discordancy of k suspicious upper outliers $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$ in a normal sample, suppose that the underlying distribution of the sample is a normal distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2 , where μ and variance σ^2 are unknown.

Under a *null hypothesis*, the sample data $x_1, x_2, \dots, x_n$ are a random sample from $N(\mu, \sigma^2)$. Under an *alternative hypothesis*, unsuspicious observations $x_1, x_2, \dots, x_{(n-k)}$ belong to $N(\mu, \sigma^2)$, but the suspicious upper outliers $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$ belonging to $N(\mu + a, \sigma^2)$ ($a > 0$), which has a larger mean $\mu + a$ shifted right from the original mean μ .

In other words, we need to test the null hypothesis

$$H_0: a = 0$$

against the mean-shifted alternative hypothesis

$$H_1: a > 0.$$

The likelihood-ratio statistic for testing $H_0: a = 0$ against $H_1: a > 0$ is

$$[x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k \cdot \bar{x}] / s,$$

where $\bar{x}$ and s stand for the sample mean and sample standard deviation, and large values of the test statistic reject the null hypothesis H_0 , identifying $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$ as outliers or discordant observations.

If the underlying distribution of the sample is a nonnormal distribution, say, a gamma distribution (which includes exponential distribution as a special case), then the likelihood-ratio test statistic will be

$$[x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)}] / \sum x_i.$$

Dixon Test

There are many varieties of statistical tests for detecting outliers in the sample. Generally speaking, powerful tests are based on sophisticated test statistics, such as the likelihood-ratio test statistics discussed before.

Dixon test is a very simple test that is often used to test outliers in a small sample. The general form of Dixon statistic in the literature is defined by

$$T = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}}, \quad p \leq r < s \leq q,$$

in terms of the ratio of intervals of ordered values.

The advantage of Dixon test is the simplicity of its test statistic for hand calculation. Because Dixon statistic simply uses the sample information from four observations $x_{(p)}$, $x_{(r)}$, $x_{(s)}$, $x_{(q)}$, the power of Dixon test is relatively low unless the sample size is very small.

There are many different versions of Dixon tests. In the simplest case,

$$T_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ or } \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$

can be used to test a single suspicious upper outlier $x_{(n)}$ or lower outlier $x_{(1)}$.

Large values of T_1 will reject the null hypothesis, identifying $x_{(n)}$ or $x_{(1)}$ as outlier. Similarly, we can use

$$T_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} \text{ or } \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$

to simultaneously test a pair of upper outliers $x_{(n-1)}$, $x_{(n)}$ or lower outliers $x_{(1)}$, $x_{(2)}$.

statistics T_1 and T_2 , assuming that the underlying distribution of the sample is normal. More detailed tables can be found in the references listed in the Further Reading section at the end of this entry.

When T_1 is larger than its critical value t_{α} , we will reject the null hypothesis at significance level α , identifying $x_{(n)}$ or $x_{(1)}$ as outlier.

Similarly, if T_2 is larger than its critical value t_{α} , we will reject the null hypothesis at significance level α , simultaneously identifying $x_{(n-1)}$, $x_{(n)}$ or $x_{(1)}$, $x_{(2)}$ as outliers.

A simple way to test multiple outliers is to apply sequential or consecutive tests. That is, sequentially apply a test for a single outlier. Nevertheless, sequential tests are generally less powerful.

For example, to test the discordancy of three suspicious upper outliers $x_{(n)}$, $x_{(n-1)}$, $x_{(n-2)}$ in the sample data, we can use an inward or outward method to test them sequentially. The inward method tests $x_{(n)}$ first. If it is accordant, stop the procedure and declare $x_{(n-1)}$ and $x_{(n-2)}$ accordant as well. Otherwise, delete $x_{(n)}$ from the sample and then repeat the same procedure to test $x_{(n-1)}$ and $x_{(n-2)}$. On the other hand, the outward method tests $x_{(n-2)}$ first by deleting $x_{(n-1)}$ and $x_{(n)}$. If $x_{(n-2)}$ is discordant, stop the procedure and declare $x_{(n-1)}$ and $x_{(n)}$

Applying Dixon Test

To apply Dixon test for outliers at a given significance level α ($0 < \alpha < 1$), we need to know the critical value or percentage point of the test statistic. If the Dixon statistic T is larger than its critical value t_{α} , reject the null hypothesis at level α , identifying the suspicious observations as outliers.

The critical value of a test statistic is determined by the sampling distribution of the statistic under the null hypothesis, which in turn depends on the underlying distribution of the sample. For example, Table 1 gives some of the critical values of Dixon

Table 1 Critical Values t_{α} of Dixon Statistics for Normal Samples

	$T_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ or } \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$		$T_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} \text{ or } \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}$	
n	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$
5	0.642	0.780	0.845	0.929
6	0.560	0.698	0.736	0.836
7	0.507	0.637	0.661	0.778
8	0.468	0.590	0.607	0.710
9	0.437	0.555	0.565	0.667
10	0.412	0.527	0.531	0.632
12	0.376	0.482	0.481	0.579
14	0.349	0.450	0.445	0.538
16	0.329	0.426	0.418	0.508
18	0.313	0.407	0.397	0.484
20	0.300	0.391	0.372	0.464
25	0.277	0.362	0.343	0.428
30	0.260	0.341	0.322	0.402

discordant as well. Otherwise, add $x_{(n-1)}$ into the sample and then test $x_{(n-1)}$, and so on.

Both inward and outward methods have advantages and disadvantages. They may suffer from masking or swamping effect, which is the inability to correctly identify an outlier in the presence of other outliers. This is always a tough issue in multiple-outlier situations.

Example

In a comparison of strength of various plastic materials, one important characteristic is the percent elongation at break. The following data are 10 measurements of percent elongation at break made on certain material:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
2.02	2.22	3.04	3.23	3.59	3.73	3.94	4.05	4.11	4.13

where $x_{(1)}$ and $x_{(2)}$ appear to be lower outliers in the sample data.

Because Dixon statistic

$$T_2 = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}} = \frac{3.04 - 2.02}{4.13 - 2.02} = 0.4834,$$

which is smaller than its 5% critical value 0.531 in Table 1 with $n = 10$, we fail to identify $x_{(1)}$ and $x_{(2)}$ as outliers at the 5% significance level.

Similarly, we will get the same result if sequential tests are used. In fact,

$$\begin{aligned} T_1 &= \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \\ &= \frac{2.22 - 2.02}{4.13 - 2.02} = 0.0948 \quad (n = 10) \end{aligned}$$

$$\begin{aligned} T_1 &= \frac{x_{(3)} - x_{(2)}}{x_{(n)} - x_{(2)}} \\ &= \frac{3.04 - 2.22}{4.13 - 2.22} = 0.4293 \quad (n = 9, \text{ deleting } x_{(1)}) \end{aligned}$$

which are respectively smaller than the 5% critical values 0.412 and 0.437. Therefore, both inward and

outward methods will fail to identify $x_{(1)}$ and $x_{(2)}$ as outliers at the 5% significance level. However, it should be pointed out that the likelihood-ratio test $[\bar{x} - x_{(1)} - x_{(2)}]/s$ will detect $x_{(1)}$ and $x_{(2)}$ as outliers at the 5% significance level.

—Jin Zhang

See also Normal Curve

Further Reading

- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.
- Dixon, W. J. (1951). Ratios involving extreme values. *Annals of Mathematical Statistics*, 22(1), 68–78.
- Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika*, 11, 46–54.
- Zhang, J. (1998). Tests for multiple upper or lower outliers in an exponential sample. *Journal of Applied Statistics*, 25, 245–255.
- Zhang, J., & Yu, K. (2004). The null distribution of the likelihood-ratio test for two upper outliers in a gamma sample. *Journal of Statistical Computation and Simulation*, 74, 461–467.
- Zhang, J., & Yu, K. (2006). The null distribution of the likelihood-ratio test for one or two outliers in a normal sample. *Test*, 15, 141–150.

Dixon test for outliers has been implemented in the R project, a free software environment for statistical computing and graphics: <http://finzi.psych.upenn.edu/R/library/outliers/html/dixon.test.html>

DUNN'S MULTIPLE COMPARISON TEST

Olive Jean Dunn's work was one of the earliest attempts to provide researchers with a way to select, in advance, and test a number of contrasts from among a set of mean scores. Fisher, Scheffé, and Tukey had already provided techniques for testing comparisons between all possible linear contrasts among a set of normally distributed variables. Dunn's contribution meant that researchers no longer needed to test all possible comparisons when they were interested in only a few such comparisons, yet they maintained control over an inflated Type I error rate.

Although the groundwork for Dunn's multiple comparison tests is usually attributed to Carlos Emilio Bonferroni, it actually originated with George Boole, who worked in the middle of the 19th century. Boole's inequality (also known as the union bound) states that for any finite set of events, the probability that at least one of the events will occur is no greater than the sum of the probabilities of the individual events. Bonferroni expanded Boole's inequality by demonstrating how upper and lower bounds (i.e., a confidence interval) could be calculated for the probability of the finite union of events. These are called Bonferroni's inequalities.

Dunn, and later, Dunn and Massey, used a Bonferroni inequality to construct simultaneous confidence intervals for k means, m comparisons, and v degrees of freedom based on the Student's t statistic. She demonstrated the differences in confidence intervals obtained when the variances of the means were unknown, and when the variances were unknown but equal; she also showed how her confidence intervals could be used in fitting data to locate regression curves (e.g., growth in height or weight). As such, no comprehensive table for different numbers of means, comparisons, and degrees of freedom was produced until B. J. R. Bailey did so in 1977. Bailey noted that Dunn's tables were incomplete, were rounded to two decimal places, and contained errors in the tabled values. Although Dunn conducted the initial work showing how complete tables might be constructed, Bailey honored the forerunner by titling his paper "Tables of the Bonferroni t Statistic"; nevertheless, the overlapping t values are, except for rounding errors, identical. To date, there remains confusion about the attribution of this multiple comparison method, no doubt partly because Bonferroni's publications were written in Italian.

Perhaps adding to the confusion, Zbynek Sidák constructed a partial set of tables using the multiplicative inequality to control family-wise Type I error, whereas Dunn had employed the additive inequality for the same purpose. Sidák showed that using the multiplicative inequality produced slightly smaller confidence intervals than using the additive inequality.

This increases the probability of finding statistically significant differences between pairs of means, making the test more powerful. Ten years later, Paul Games published a more complete set of tables using Sidák's method. Nowadays, one often sees references to the Dunn-Sidák multiple comparison test, but, as noted above, the two methods are not identical and produce somewhat different results.

Why the Dunn Multiple Comparison Test Is Used

Dunn's multiple comparison test is an adjustment used when several comparisons are performed simultaneously. Although a value of alpha may be appropriate for one individual comparison, it is not appropriate for the set of all comparisons. In order to avoid a surfeit of Type I errors, alpha should be lowered to account for the number of comparisons tested.

Suppose a researcher has collected data on 20 independent variables (e.g., gender, intact vs. divorced parents, age, family income, etc.) that might (or might not) be related to some dependent variable of interest (e.g., level of physical violence displayed). The researcher might be tempted to go on a fishing expedition, making comparisons between all possible means produced by each independent variable (e.g., boys vs. girls, rich vs. poor families, etc.) In all, 20 comparisons could be made. If the alpha level was set at .05, the researcher has a pretty good chance of finding at least one statistically significant difference even if all of the independent variables are completely unrelated to displays of physical violence. That is because when the alpha level is set at .05, we know that the odds of obtaining a difference deemed statistically significant would happen by chance on only 1 out of 20 occasions on repeated sampling. In this example, the researcher has given him- or herself 20 chances of finding a statistically significant difference, so it is not surprising that one of them met the critical value for significance. The Dunn adjustment effectively raises the standard of evidence needed when researchers are comparing a large number of means simultaneously.

How to Use the Dunn Multiple Comparison Test

The simplest way to understand the Dunn procedure is to understand that, if we are making n comparisons instead of just one, we must divide the selected alpha by n . For example, if we were testing the effect of social praise on the math achievement of elementary, middle school, and high school students, we should not set alpha at the traditional .05 level, but at the $\alpha = .05/3$, or .0167 level. This ensures that across all three comparisons, the chance of making a Type I error remains at .05.

Dunn's approach in creating confidence intervals is simply a different way of accomplishing the same outcome. Table 1 represents Dunn's multiple comparisons for the effect of three levels of curiosity on externalizing affect scores. To obtain the confidence interval for the comparison of high versus medium curiosity, simply multiply the standard error of the difference (SE_{diff}) for the contrast by Dunn's tabled t value: $1.52 \times 2.40 = 3.648$, or 3.65. This is the value that must be subtracted from and added to the mean difference (Ψ) to obtain the lower and upper limit for the 95% confidence interval: $\Psi \pm 3.65$. Therefore, $10.41 - 3.65 = 6.76$, and $10.41 + 3.65 = 14.06$. The odds are 95 out of 100 that the true difference between the means for high curiosity and medium curiosity is contained within the confidence interval from 6.76 to 14.06 (i.e., $6.76 \leq \Psi \leq 14.06$). Ninety-five percent confidence intervals that do not contain the value zero are considered statistically significant. Because none of the three confidence intervals contains zero, the researcher would have a high degree of confidence in saying that each level of curiosity produced a mean that was significantly different from every other level, with high curiosity leading to the highest level of externalizing affect, and low curiosity leading to the lowest level of externalizing affect. The probabilities in Table 1 were adjusted for the fact that multiple comparisons were made.

Table 1 Dunn Multiple Comparisons for the Effect of Three Levels of Curiosity on Externalizing Affect Scores

Comparison	df	Mean Difference	SE_{diff}	t	p	95% Confidence Interval
Curiosity High vs. Medium	703	10.41	1.52	2.40	.000	6.76 to 14.06
High vs. Low	456	14.34	1.77	2.40	.000	10.09 to 18.59
Medium vs. Low	702	3.92	1.52	2.40	.003	0.27 to 7.57

Final Considerations

When the variables in a comparison are correlated, the normal Dunn correction is more conservative, causing less power, so further adjustment should be used. In this case, the corrected alpha falls between the usual Dunn correction and no correction at all.

The decision to use (or not use) a multiple comparison test like Dunn's hinges not on the fact that several comparisons are to be made, but on an understanding of the theory and logic of the research design. For instance, an investigator may collect data to test three independently derived hypotheses, say, the relationship between gender and anxiety, the influence of three kinds of primary reinforcer on reading achievement, and the impact of physical size (small and large) on the extent of physical violence displayed. Although five comparisons can be made with these data, only the three hypotheses associated with the primary reinforcer data should be corrected with Dunn's multiple comparison test.

—Ronald C. Eaves and Anthony J. Guarino

See also Bonferroni Test; Post Hoc Comparisons; Tukey-Kramer Procedure

Further Reading

Bailey, B. J. R. (1977). Tables of the Bonferroni t statistic. *Journal of the American Statistical Association*, 72, 469–478.

- Dunn, O. L. (1961). Multiple comparisons of means. *Journal of the American Statistical Association*, 56, 52–64.
- Dunn, O. L., & Massey, F. J., Jr. (1965). Estimation of multiple contrasts using t -distributions. *Journal of the American Statistical Association*, 60, 573–583.
- Games, P. A. (1977). An improved t table for simultaneous control on g contrasts. *Journal of the American Statistical Association*, 72, 531–534.
- Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. London: Chapman & Hall.
- Sidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626–633.
- Bonferroni correction: <http://www.cmh.edu/stats/ask/bonferroni.asp>
- Bonferroni correction/adjustment: <http://home.clara.net/sisa/bonhlp.htm>
- Boole's inequality and Bonferroni's inequalities description: http://www.absoluteastronomy.com/encyclopedia/b/bo/booles_inequality.htm

E

Behavioral psychology is the science of pulling habits out of rats.

—Dr. Douglas Busch

ECOLOGICAL MOMENTARY ASSESSMENT

Ecological momentary assessment (EMA) allows the study of behavior, psychological states, and physiological functions in their natural contexts. EMA and its predecessors (e.g., the experience sampling method) were developed with several purposes in mind. First, there was concern that retrospective autobiographical memory was fallible, due primarily to the use of cognitive heuristics during recall. EMA reduces these biases by generally limiting the period over which information is recalled. Second, although laboratory research offers the benefit of experimental control, it is not clear if processes observed in the laboratory are similar to what occurs in the “real” world. EMA often has greater ecological validity and generalizability because assessments can be collected in everyday settings. Third, EMA enables a closer examination of dynamic and temporal processes. EMA designs usually incorporate a large number of repeated measures, which provide a movielike view of processes over time. Such data not only allow examination of temporal patterns but also provide considerable information about (although not confirming) causal associations among variables.

EMA involves participants reporting on current or recent psychological states, behaviors, and/or environmental conditions, typically multiple times each day, for days or even weeks. Responses are collected in several ways, of which some are self-initiated by research participants and others request responses after some signal (e.g., a pager, a handheld computer alarm). The three most commonly used approaches are interval-contingent, event-contingent, and signal-contingent responding. *Interval-contingent recording* involves completing assessments at regular times (e.g., every hour on the hour, before bed). *Event-contingent schedules* entail completing assessments in response to specific events (e.g., smoking a cigarette, argument with a spouse). *Signal-contingent schedules* require individuals to report on experiences in response to random or semirandom signals across the day. Recent technological advances, most notably palmtop computers, provide a number of advantages to EMA data capture over paper-and-pencil approaches. As participants can respond directly on a handheld computer, portability is optimized, compliance can be automatically tracked (reports are date- and time-stamped), data can be transferred directly to statistical software, and researchers have greater control over the format and order of assessment items.

Despite the advantages of EMA approaches, they are not without limitations. First, implementation of EMA designs requires considerable time and expertise. There are many logistical issues: the design of the sampling scheme, thoughtful consideration of questionnaire design, training and motivating participants to follow the protocol, and dealing with the technical difficulties inherent in the use of technological devices (e.g., programming the devices). Second, momentary data collection techniques yield masses of complex, time-dependent data. Although such data are a strength of the approach, considerable statistical and data management acumen are necessary to manipulate and appropriately analyze these data sets. Third, given the intensive nature of data collection (e.g., five times each day for 2 weeks), the majority of participants are likely to have some missing data. This presents a problem for EMA research and must be accounted for in the statistical/analytic approach and interpretation of the data (e.g., are missing data random or reflective of an altered environmental state?).

Conclusion

EMA and other strategies for capturing momentary data provide researchers with a new assessment technique for studying behavior, psychological states, and physiological functions as they occur in individuals' natural environments. This method can reduce retrospective recall biases, provides a dynamic picture of people's daily lives, and may reveal potential causal relationships among variables of interest. New technological advances, such as palmtop computers and interactive voice recognition systems, are opening up exciting new avenues for real-time data capture in naturalistic settings.

—Joshua Smyth and Kristin Heron

See also Data Mining

Further Reading

Stone, A., & Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, 16, 199–202.

Experience sampling method: <http://seattleweb.intel-research.net/projects/ESM/index.html>

EDUCATIONAL TESTING SERVICE

Educational Testing Service (ETS) is the world's largest private educational testing and measurement organization. With an annual budget approaching \$1 billion, it develops, administers, or scores more than 24 million tests annually (as of 2005) in more than 180 countries at more than 9,000 locations internationally. With locations worldwide, its operations are headquartered in Princeton, New Jersey.

ETS was founded by Henry Chauncey in 1947, with key support from the American Council on Education, The Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board. The core ideas behind ETS were put forth by former Harvard president James Conant. Its mission is to “advance quality and equity in education for all people worldwide.”

ETS encompasses five areas—research, assessment development, test administration, test scoring, and instructional products and services—but it is best known for assessment development. Perhaps its most well-known test, the SAT®, is actually published by the College Board, although ETS develops and administers the test as a work-for-hire (a procedure it also does for the Advanced Placement Exams). The SAT I measures Mathematics, Critical Reading, and Writing. The SAT IIs are subject-based tests that assess particular areas of learning.

Major ETS assessments include the GRE® (Graduate Record Examinations), TOEFL® (Test of English as a Foreign Language), and the Praxis tests for teacher certification. The GRE has three subtests: Verbal, Quantitative, and Analytical Writing. The latter subtest replaced an Analytic Reasoning subtest on October 1, 2002. The TOEFL exam is currently paper based, although there is a shift toward an Internet-based measure. The paper-based measure assesses Listening Comprehension, Structure and Written Expression, and Reading Comprehension. The new Internet-based measure assesses along the four dimensions of Listening, Structure, Reading, and Writing. The Praxis is a series of three different tests: the first measures basic academic skills; the second measures general and subject-specific

knowledge and teaching skills; and the third measures classroom performance. These three assessments represent only a hint of the many different ones offered by ETS.

ETS has many critics, perhaps most notably FairTest (although such critics tend to be against all methods of standardized testing, not solely ETS). FairTest and other critics argue that ETS tests are biased, overly coachable, and prone to misuse. One example of such criticisms is that men outperform women on the GRE Quantitative test by nearly a standard deviation despite the fact that women often outperform men in advanced mathematics in the classroom. The performance-based assessments that FairTest advocates, however, would likely create similar confounding issues.

Perhaps the primary competitor of ETS is the ACT, formerly the American College Testing Program, which produces the ACT tests. The ACT test is usually accepted in lieu of the SATs by most universities. In addition, ACT recently took over the development of the GMAT (Graduate Management Admissions Test) from ETS.

—James C. Kaufman

See also Ethical Issues in Testing; Standards for Educational and Psychological Testing

Further Reading

Educational Testing Service: <http://www.ets.org>

FairTest: <http://www.fairtest.org>

EDWARDS PERSONAL PREFERENCE SCHEDULE

The *Edwards Personal Preference Schedule* (EPPS) (publisher: The Psychological Corporation) is a scale designed to measure 15 personal needs, originally proposed by H. A. Murray. The scale, authored by Allen Edwards, was constructed to provide ipsative information on how people rank one need relative to their other needs, as well as normative information on their needs compared with other people's. Edwards discussed needs as nonclinical personality variables

and considers the EPPS foremost a personality measure. The EPPS has been used in vocational counseling to encourage discussion about how individuals want to relate to coworkers and their desired levels of responsibility on the job.

The EPPS includes 15 personality scales and two scales for assessing the validity of an individual's results. The personality dimensions include needs for Achievement (succeeding and fulfilling high standards), Deference (concern for the opinions of or approval from others), Order (organization and fastidiousness), Exhibition (social attention), Autonomy (freedom to self-determine), Affiliation (attachment to friends), Intracception (psychological-mindedness and introspection), Succorance (sympathy and affection from others), Dominance (leading and decision making), Abasement (feeling guilt for wrongdoings), Nurturance (helping others), Change (variety and novelty of activity), Endurance (task focus and forbearance), Heterosexuality (engaging the opposite sex romantically or sexually), and Aggression (being derisive, critical, and vengeful toward others).

The items on the EPPS pair two statements, each reflecting 1 of the 15 dimensions, and require test takers to identify which is more typical of themselves. Statements reflecting each of the personality variables are paired two times with statements reflecting each of the others. Overall, the test requires approximately 45 minutes to administer. Raw scores from the test can be used to identify the relative importance of a need to an individual, whereas normative data, collected in the late 1950s from a college sample and a survey of U.S. households, supplies information on how test takers' personal needs compare with others'.

The test and its norms were most recently updated in 1959; consequently, the instrument has been criticized for having normative data that are too old to serve as a meaningful index. However, the EPPS has been praised for the degree to which its item structure reduces the influence of social desirability, and overall evidence suggests that the individual scales show moderate-to-favorable internal consistency and satisfactory stability over a week. In addition, the evidence of convergent validity for the scale scores, most of which was collected in the 1950s, shows that

the EPPS scale scores relate modestly, though as predicted, to other personality measures. For example, measures of agreeableness showed small positive correlations with deference and nurturance scales and small-to-medium negative correlations with aggression and dominance scales.

—Matthew E. Kaler and Jo-Ida C. Hansen

See also Jackson Personality Inventory–Revised; Personality Tests

Further Reading

- Edwards, A. L. (1959). *Manual: Edwards Personal Preference Schedule*. Washington, DC: The Psychological Corporation.
- Helms, J. E. (1983). *Practitioner's guide to the Edwards Personal Preference Schedule*. Springfield, IL: Charles C Thomas.
- Thorson, J. A., & Powell, F. C. (1992). Vagaries of college norms for the Edwards Personal Preference Schedule. *Psychological Reports*, 70, 943–946.

EFFECT SIZE

Effect size is a term used to describe the magnitude of a treatment effect. More formally, it can be defined as the degree to which the null hypothesis is false, versus a true alternative hypothesis. Measuring effect size has taken on increasing study and importance during the last 30 years. Effect size is important in three phases of the research process. First, it is important prior to collecting data, as it is required for estimating sample sizes that are necessary to ensure statistical power. Second, reporting effect size is important for interpreting statistical tests of significance. Reporting effect size is fundamental to any good statistical report of results, since statistical tests and their associated p values are functions of both effect size and sample size. Finally, effect size measures are the raw scores of a meta-analysis.

Effect size measures attempt to strip away the effects of sample size and produce a simple and easily interpretable measure of the size of the effect. To achieve this goal, effect size indices have the following three properties: (1) standardization to allow

cross-study comparison and to enhance interpretability of unfamiliar scales, (2) preservation of the direction of the effect, and (3) independence from sample size. Three common classes of effect size measures are standardized mean differences, correlation coefficients or their squared values, and odds ratios (common in medical but not educational and behavioral research). For many common statistical tests (e.g., t tests), the raw or unstandardized effect size is the numerator of the test statistic. The use of the “unstandardized” mean difference along with a confidence interval may be preferable to reporting effect size statistics when the dependent variable is measured on a familiar scale or is readily understood in the field of study.

Why It Is Important to Report Effect Size

Statistical tests combine in various ways effect size and sample size. Equation 1 summarizes the general relationship between tests of statistical significance and effect size. Clearly, the value of the statistical test and its corresponding p value depend on both effect size and sample size. Reexpressing the equation shows that by dividing a statistical test by its sample size, it is possible to get a measure of the effect that is independent of sample size:

$$\begin{aligned} \text{Significance test} &= \text{Effect size} \times \text{Sample size or} \\ \text{Effect size} &= \frac{\text{Significance test}}{\text{Sample size}}. \end{aligned} \quad (1)$$

In 1991, Rosenthal provided a comprehensive set of formulas that describe the specific relationship between tests of significance, effect size, and study sample size. Table 1 shows this relationship for a few important statistical tests.

Reporting effect size helps avoid the misleading decision dichotomy ($p \leq .05$ versus $p > .05$) inherent in classical null hypothesis statistical testing by “stripping away” the effect of sample size from the test statistic. With the sample size factored out, measures of effect size help researchers answer the

Table 1 Selected Statistics: Statistic¹ = Effect Size × Study Size

<i>Chi-Square Test for Contingency Table</i>	<i>Independent Sample t Test</i>	<i>Two Independent Sample F Test</i>
$\chi^2 = \phi^2 \times N$	$t = [(\bar{X}_1 - \bar{X}_2)/S_{(\text{pooled})}] \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = d \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$F = [(\bar{X}_1 - \bar{X}_2)/S^2] \times \frac{n_1 n_2}{n_1 + n_2}$

1. Note that because a *p* value is determined by the value of the test statistic, it is also a function of the effect size and study size.

important clinical or practical question: What is the size of the treatment or intervention effect? The publication style manuals for both the American Psychological Association (2001) and the American Medical Association, among others, require that measures of effect size be reported with statistical tests.

Measuring Effect Size: Two Types of Effect Size Indices

There are basically two types of effect size measures: measures of standardized or relative mean differences and measures of relationship or correlation, including measures of relationship for nonquantitative variables. Most measures of effect size overestimate the population effect size. However, because overestimation is small except when sample size is extremely small, adjustments for bias are seldom used. Table 2 shows these two types of effect size measures for a few common statistics. Because statistical significance is monotonically related to sample size, it is important to report effect size for both statistically significant and nonsignificant results. Presently, there is little consensus regarding which of the various effect size measures to report. Wolf, in 1986, gave available formulas for converting various test statistics (X^2 , t , F) to an effect size measure, and computer programs for these calculations are also readily available (see Computer Programs for Calculating Effect Size). A comprehensive list of effect size measures, their formulas, and calculation examples is found in Rosenthal (for a focus on control versus treatment effect size measures); Grissom and Kim (for a

comprehensive review of effect size measures, especially when violations to the normal model are suspected); and Olejnik and Algina (for an overview of measures of effect size for multigroup comparative studies).

Standardized Mean Difference

Cohen's d is a measure of effect size in standard deviation units that is very commonly used as a basis to estimate the sample sizes required to ensure statistical power for a two-sample problem and as the data for conducting a meta-analysis. To calculate d , positive values are assigned to mean differences favoring the treatment, and negative values are assigned to differences unfavorable to the treatment. The scale used to standardize the mean difference is, alternatively, the control group alone (Glass's g' index, preferred when the treatment affects the variance) or for two repeated measures, the pretest (using the pretest ensures that the treatment does not affect the variance); the pooled standard deviation of the groups that are compared (Cohen's d , preferred when the treatment does not affect the variance in important ways); or the pooled standard deviation for all of the treatments in multigroup designs (Hedges's g).

Correlation and Proportion of Variance Measures

The point-biserial correlation coefficient (r_{pb}), for t tests for two independent samples; the phi coefficient (ϕ), for chi-square tests for contingency tables

Table 2 Effect Size Measures for Selected Test Statistics

Test Statistic	Effect Measure		
	Raw Unstandardized Effect ² (Mean Difference)	Relative Effect ¹ / RawEffect/Standard Error or <i>t</i> /Sample Size Ratio	Correlation or Correlation Squared
One-Sample <i>t</i>	$(\bar{X}_1 - \mu)$	$d = [(\bar{X}_1 - \mu)/S_x]$ $d = \frac{t}{\sqrt{df}}$	$r = \sqrt{\frac{t^2}{t^2 + df}}$
Two-Sample <i>t</i>	$(\bar{X}_1 - \bar{X}_2)$	$d = [(\bar{X}_1 - \bar{X}_2)/S_{(pooled)}]$ $d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	$r^2 = \eta^2 = \frac{t^2}{t^2 + df}$ or $\omega^2 = \frac{t^2 - 1}{t^2 + df + 1}$
Dependent Sample <i>t</i>	$(\bar{D} - \mu)$	$d = [(\bar{D} - \mu)/S_D]$ $d = \frac{2t}{\sqrt{df}}$	
One-Way ANOVA ³	$(\bar{X}_{\text{largest}} - \bar{X}_{\text{smallest}})$ or MS_A	$f_{\text{effect}} = \frac{S_A}{S_e} = \frac{\sqrt{MS_A}}{\sqrt{MS_e}}$	$\omega_A^2 = \frac{(k-1)(F-1)}{(k-1)(F-1) + kn}$ $r^2 = \eta^2 = \frac{F}{F + df_{\text{error}}} = \frac{SS_A}{SS_{TO}}$

1. Cohen's *d* expresses the raw effect size relative to the standard deviation.
2. The raw effect serves as the numerator for many standard tests for mean differences.
3. *f* is an extension of *d* to ANOVA and multigroup designs and is used to calculate sample size estimates (Cohen, 1988, effect size interpretive guidelines: small = .10; medium = .25; large = .40).

when both variables are dichotomous; and Spearman's rank order correlation coefficient, when both variables are rank ordered, are all common measures of effect size.

Measures of the proportion of variance in the dependent variable that is associated with the independent variable, such as the point-biserial correlation coefficient squared (r^2 , n^2 , and ω^2), are all used to report effect size. Eta squared and partial eta squared are estimates of the degree of association for the sample. Omega squared, virtually almost always smaller than

eta squared, estimates the degree of association in the population. SPSS displays eta or partial eta squared when the "display effect size" option is selected in the program GLM. A limitation of squared values is that they can obscure the direction of the effect.

Interpreting Effect Size Measures

Table 3 summarizes widely used guidelines proposed by Cohen for the interpretation of effect size in the

behavioral and social sciences. Although widely used, these guidelines suffer from a number of limitations. Interpretation of an effect size depends on a number of factors, including the specific research design used, knowledge of the substantive area of study, an appreciation of real-world consequences, and the theoretical importance of the effect. For example, effect sizes for independent and dependent group designs may be difficult to compare, as they control for different sources of bias. A small effect that is easy to implement, does not have adverse consequences or side effects, and requires few resources may provide small but important benefits. Small effects may be important for serious and difficult-to-treat problems. Small effects when applied to an individual may be important when applied to the entire population. For example, following these guidelines, many biomedically useful treatments have very small effects (cf. Salk vaccine and paralytic polio, $r = .01$; and psychotherapy and improvement, $r = .39$). Ultimately, interpretation of the effect size is extrastatistical and based on the knowledge of the field of study. Some argue that because importance is ultimately not a statistical issue, reporting effect size measures serves only to obfuscate further understanding of the results.

Rosnow and Rosenthal have argued that proportion-of-variance measures are prone to misinterpretation and encourage an overly pessimistic interpretation of many “small” but important effects. Instead, they have suggested that researchers report measures of correlation. For example, some suggest that r be

interpreted in terms of success rates (e.g., Rosenthal & Rubin’s binomial effect size display [BESD]) in the treatment and comparison groups, assuming an overall success rate of 50%. Mathematically, the BESD is a transformation of r to X^2 , where success rate = $0.50 \pm r/2$. For example, the medium effect size $r = .30$ [$r^2 = .09$] converts to a BESD comparison group success rate of .35 and to a treatment group success rate of .65. For an alternative intuitive measure, see Cohen’s $U3$, which describes the percentage of scores in the comparison group that was exceeded by the mean score in the treatment group.

Other factors that influence the effect size include the range of treatments studied (increasing the range of treatments generally increases effect size). Almost any violation of the normal model may cloud interpretation. Nonnormality, heterogeneity of variance, low measurement reliability, and the presence of outliers all influence effect size estimation. In short, interpreting effect size and comparing effect sizes across studies requires attention to the specific design features of the various studies and the assumptions underlying the statistical tests. In light of these limitations, Table 4 summarizes perspectives for interpreting effect size measures.

Computer Programs for Calculating Effect Size

Effect Size Calculator is an especially user-friendly shareware program. Effect Size Determination Program is somewhat more comprehensive and designed to facilitate coding for a meta-analysis. Most meta-analysis computer programs calculate one or more effect size measures (e.g., MetaWin).

—Ward Rodriguez

Table 3 Cohen’s Guidelines for Interpreting Effect Size Measures

Effect Size Measure	Small	Medium	Large
Standardized Mean Difference	0.20	0.50	0.80
Correlation Coefficient	0.10	0.30	0.50
Correlation Squared	0.01	0.06	0.14

See also Significance Level; Type I Error; Type II Error

Further Reading

- Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Erlbaum.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Table 4 Perspectives for Interpreting Effect Size Measures

Cohen's Guidelines	d	r	r^2	Percentile Standing	Percent Nonoverlap Control vs. Treatment
	2	0.707	0.5	97.7	81.10%
	1.9	0.689	0.474	97.1	79.40%
	1.8	0.669	0.448	96.4	77.40%
	1.7	0.648	0.419	95.5	75.40%
	1.6	0.625	0.39	94.5	73.10%
	1.5	0.6	0.36	93.3	70.70%
	1.4	0.573	0.329	91.9	68.10%
	1.3	0.545	0.297	90	65.30%
	1.2	0.514	0.265	88	62.20%
	1.1	0.482	0.232	86	58.90%
	1	0.447	0.2	84	55.40%
Large	0.9	0.401	0.168	82	51.60%
	0.8	0.371	0.138	79	47.40%
	0.7	0.33	0.109	76	43.00%
Medium	0.6	0.287	0.083	73	38.20%
	0.5	0.243	0.059	69	33.00%
	0.4	0.196	0.038	66	27.40%
Small	0.3	0.148	0.022	62	21.30%
	0.2	0.1	0.01	58	14.70%
	0.1	0.05	0.002	54	7.70%
	0	0	0	50	0%

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Hillsdale, NJ: Erlbaum.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241–286.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

EIGENDECOMPOSITION

Eigenvectors and *eigenvalues* are numbers and vectors associated with square matrices, and together they provide the *eigendecomposition* of a matrix, which

analyzes the structure of this matrix. Even though the eigendecomposition does not exist for all square matrices, it has a particularly simple expression for a class of matrices often used in multivariate analysis, such as correlation, covariance, or cross-product matrices. The eigendecomposition of this type of matrices is important in statistics because it is used to find the maximum (or minimum) of functions involving these matrices. For example, principal component analysis is obtained from the eigendecomposition of a covariance matrix and gives the least square estimate of the original data matrix.

Eigenvectors and eigenvalues are also referred to as *characteristic vectors* and *latent roots* or *characteristic equation* (in German, *eigen* means “specific to” or “characteristic of”). The set of eigenvalues of a matrix is also called its *spectrum*.

Notations and Definition

There are several ways to define eigenvectors and eigenvalues. The most common approach defines an eigenvector of the matrix $\mathbf{A}$ as a vector $\mathbf{u}$ that satisfies the following equation:

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}. \quad (1)$$

When rewritten, the equation becomes

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}, \quad (2)$$

where λ is a scalar called the eigenvalue associated to the eigenvector.

In a similar manner, a vector $\mathbf{u}$ is an eigenvector of a matrix $\mathbf{A}$ if the length of the vector (but not its direction) is changed when it is multiplied by $\mathbf{A}$. For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \quad (3)$$

has the eigenvectors

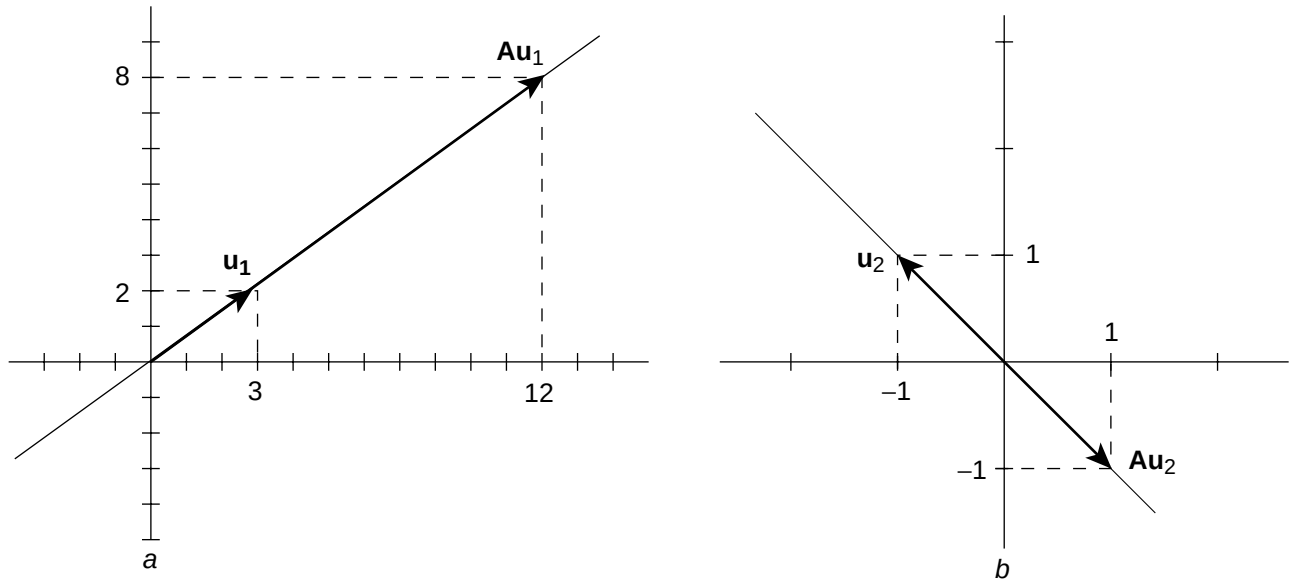


Figure 1 Two Eigenvectors of a Matrix

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \text{with eigenvalue } \lambda_1 = 4 \quad (4)$$

and

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{with eigenvalue } \lambda_2 = -1. \quad (5)$$

We can verify (as illustrated in Figure 1) that only the lengths of $\mathbf{u}_1$ and $\mathbf{u}_2$ are changed when one of these two vectors is multiplied by the matrix $\mathbf{A}$:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 4 \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} \quad (6)$$

and

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \quad (7)$$

For most applications, we normalize the eigenvectors (i.e., transform them so that their length is equal to one):

$$\mathbf{u}^T \mathbf{u} = 1. \quad (8)$$

For the previous example, we obtain

$$\mathbf{u}_1 = \begin{bmatrix} .8331 \\ .5547 \end{bmatrix}. \quad (9)$$

We can check that

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} .8331 \\ .5547 \end{bmatrix} = \begin{bmatrix} 3.3284 \\ 2.2188 \end{bmatrix} = 4 \begin{bmatrix} .8331 \\ .5547 \end{bmatrix} \quad (10)$$

and

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} -.7071 \\ .7071 \end{bmatrix} = \begin{bmatrix} .7071 \\ -.7071 \end{bmatrix} = -1 \begin{bmatrix} -.7071 \\ .7071 \end{bmatrix}. \quad (11)$$

Traditionally, we put together the set of eigenvectors of $\mathbf{A}$ in a matrix denoted $\mathbf{U}$. Each column of $\mathbf{U}$ is an eigenvector of $\mathbf{A}$. The eigenvalues are stored in a diagonal matrix (denoted $\mathbf{\Lambda}$), where the diagonal elements give the eigenvalues (and all the other values are zeros). The first equation can be rewritten as follows:

$$\mathbf{A}\mathbf{U} = \mathbf{\Lambda}\mathbf{U}; \quad (12)$$

or also as

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}. \quad (13)$$

For the previous example, we obtain

$$\begin{aligned}\mathbf{A} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} \\ &= \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ -4 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}.\end{aligned}\quad (14)$$

It is important to note that not all matrices have eigenvalues. For example, the matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ does not have eigenvalues. Even when a matrix has eigenvalues and eigenvectors, the computation of the eigenvectors and eigenvalues of a matrix requires a large number of computations and is therefore better performed by computers.

Digression: An Infinity of Eigenvectors for One Eigenvalue

It is only through a slight abuse of language that we can talk about *the* eigenvector associated with *one* given eigenvalue. Strictly speaking, there is an *infinity* of eigenvectors associated with each eigenvalue of a matrix. Because any scalar multiple of an eigenvector is still an eigenvector, there is, in fact, an (infinite) family of eigenvectors for each eigenvalue, but they are all proportional to each other. For example,

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (15)$$

is an eigenvector of the matrix $\mathbf{A}$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}. \quad (16)$$

Therefore,

$$2 \times \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} \quad (17)$$

is also an eigenvector of $\mathbf{A}$:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix} = -1 \times 2 \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \quad (18)$$

Positive (Semi-) Definite Matrices

A type of matrices used very often in statistics is called *positive semi-definite*. The eigendecomposition of these matrices always exists and has a particularly convenient form. A matrix is said to be positive semi-definite when it can be obtained as the product of a matrix by its transpose. This implies that a positive semi-definite matrix is always symmetric. So, formally, the matrix $\mathbf{A}$ is positive semi-definite if it can be obtained as

$$\mathbf{A} = \mathbf{X}\mathbf{X}^T \quad (19)$$

for a certain matrix $\mathbf{X}$ (containing real numbers). Positive semi-definite matrices of special relevance for multivariate analysis include correlation matrices, covariance matrices, and cross-product matrices.

The important properties of a positive semi-definite matrix are that its eigenvalues are always positive or null and that its eigenvectors are pairwise orthogonal when their eigenvalues are different. The eigenvectors are also composed of real values (these last two properties are a consequence of the symmetry of the matrix). Because eigenvectors corresponding to different eigenvalues are orthogonal, it is possible to store all the eigenvectors in an *orthogonal matrix* (recall that a matrix is orthogonal when the product of this matrix by its transpose is a diagonal matrix).

This implies the following equality:

$$\mathbf{U}^{-1} = \mathbf{U}^T. \quad (20)$$

We can, therefore, express the positive semi-definite matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad (21)$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ are the normalized eigenvectors; if they are not normalized, then $\mathbf{U}^T\mathbf{U}$ is a diagonal matrix.

For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad (22)$$

can be decomposed as

$$\begin{aligned} \mathbf{A} &= \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \\ &= \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \\ &= \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \end{aligned} \quad (23)$$

with

$$\begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (24)$$

Diagonalization

When a matrix is positive semi-definite, Equation 21 can be rewritten as follows:

$$\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \iff \mathbf{\Lambda} = \mathbf{U}^T \mathbf{A} \mathbf{U}. \quad (25)$$

This shows that we can transform the matrix $\mathbf{A}$ into an equivalent *diagonal matrix*. As a consequence, the eigendecomposition of a positive semi-definite matrix is often referred to as its *diagonalization*.

Another Definition for Positive Semi-Definite Matrices

A matrix $\mathbf{A}$ is said to be positive semi-definite if we observe the following relationship for any nonzero vector $\mathbf{x}$:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad \forall \mathbf{x} \quad (26)$$

(when the relationship is ≤ 0 , the matrix is *negative semi-definite*).

When all the eigenvalues of a symmetric matrix are positive, the matrix is *positive definite*. In that case, Equation 26 becomes

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad \forall \mathbf{x}. \quad (27)$$

Trace, Determinant, and Rank

The eigenvalues of a matrix are closely related to three important numbers associated to a square matrix, namely its trace, determinant, and rank.

Trace

The *trace* of a matrix $\mathbf{A}$ is denoted $\text{trace}\{\mathbf{A}\}$ and is equal to the sum of its diagonal elements. For example, with the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad (28)$$

we obtain

$$\text{trace}\{\mathbf{A}\} = 1 + 5 + 9 = 15. \quad (29)$$

The trace of a matrix is also equal to the sum of its eigenvalues,

$$\text{trace}\{\mathbf{A}\} = \sum_{\ell} \lambda_{\ell} = \text{trace}\{\mathbf{\Lambda}\}, \quad (30)$$

with $\mathbf{\Lambda}$ being the matrix of the eigenvalues of $\mathbf{A}$. For the previous example,

$$\mathbf{\Lambda} = \text{diag}\{16.1168, -1.1168, 0\}. \quad (31)$$

We can verify that

$$\text{trace}\{\mathbf{A}\} = \sum_{\ell} \lambda_{\ell} = 16.1168 + (-1.1168) = 15. \quad (32)$$

Determinant and Rank

Another classic quantity associated to a square matrix is its *determinant*. This concept of determinant, which was originally defined as a combinatoric notion,

plays an important role in computing the inverse of a matrix and in finding the solution of systems of linear equations (the term *determinant* is used because this quantity determines the existence of a solution in systems of linear equations). The determinant of a matrix is also equal to the product of its eigenvalues. Formally, if $|\mathbf{A}|$ is the determinant of $\mathbf{A}$,

$$|\mathbf{A}| = \prod_{\ell} \lambda_{\ell} \text{ with } \lambda_{\ell} \text{ being the } \ell\text{-th eigenvalue of } \mathbf{A}. \quad (33)$$

For example, the determinant of matrix $\mathbf{A}$ (from the previous section) is equal to

$$|\mathbf{A}| = 16.1168 \times -1.1168 \times 0 = 0. \quad (34)$$

Finally, the *rank* of a matrix can be defined as being the number of nonzero eigenvalues of the matrix. In this example,

$$\text{rank}\{\mathbf{A}\} = 2. \quad (35)$$

For a positive semi-definite matrix, the rank corresponds to the dimensionality of the Euclidean space, which can be used to represent the matrix. A matrix whose rank is equal to its dimensions is called *full rank*. When the rank of a matrix is smaller than its dimensions, the matrix is called *rank-deficient*, *singular*, or *multicollinear*. Only full-rank matrices have an inverse.

Statistical Properties of the Eigendecomposition

The eigendecomposition is important because it is involved in problems of optimization. For example, in principal component analysis, we want to analyze an $I \times J$ matrix $\mathbf{X}$, where the rows are observations and the columns are variables describing these observations. The goal of the analysis is to find row *factor scores*, such that these factor scores “explain” as much of the variance of $\mathbf{X}$ as possible and the sets of factor scores are pairwise orthogonal. This amounts to defining the factor score matrix as

$$\mathbf{F} = \mathbf{X}\mathbf{P}, \quad (36)$$

under the constraints that

$$\mathbf{F}^T \mathbf{F} = \mathbf{P}^T \mathbf{X}^T \mathbf{X} \mathbf{P} \quad (37)$$

is a diagonal matrix (i.e., $\mathbf{F}$ is an orthogonal matrix) and that

$$\mathbf{P}^T \mathbf{P} = \mathbf{I} \quad (38)$$

(i.e., $\mathbf{P}$ is an orthonormal matrix). There are several ways of obtaining the solution to this problem. One possible approach is to use the technique of Lagrangian multipliers, where the constraint from Equation 38 is expressed as multiplication by a diagonal matrix of Lagrangian multipliers, denoted $\mathbf{\Lambda}$, in order to give the following expression:

$$\mathbf{\Lambda}(\mathbf{P}^T \mathbf{P} - \mathbf{I}). \quad (39)$$

This amounts to defining the following equation:

$$\mathcal{L} = \mathbf{F}^T \mathbf{F} - \mathbf{\Lambda}(\mathbf{P}^T \mathbf{P} - \mathbf{I}) = \mathbf{P}^T \mathbf{X}^T \mathbf{X} \mathbf{P} - \mathbf{\Lambda}(\mathbf{P}^T \mathbf{P} - \mathbf{I}). \quad (40)$$

To find the values of $\mathbf{P}$ that give the maximum values of $\mathcal{L}$, we first compute the derivative of $\mathcal{L}$ relative to $\mathbf{P}$,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{P}} = 2\mathbf{X}^T \mathbf{X} \mathbf{P} - 2\mathbf{\Lambda} \mathbf{P}, \quad (41)$$

and then set this derivative to zero:

$$\mathbf{X}^T \mathbf{X} \mathbf{P} - \mathbf{\Lambda} \mathbf{P} = \mathbf{0} \iff \mathbf{X}^T \mathbf{X} \mathbf{P} = \mathbf{\Lambda} \mathbf{P}. \quad (42)$$

Because $\mathbf{\Lambda}$ is diagonal, this is clearly an eigendecomposition problem, and this indicates that $\mathbf{\Lambda}$ is the matrix of eigenvalues of the positive semi-definite matrix $\mathbf{X}^T \mathbf{X}$ ordered from the largest to the smallest and that $\mathbf{P}$ is the matrix of eigenvectors of $\mathbf{X}^T \mathbf{X}$ associated to $\mathbf{\Lambda}$. Finally, we find that the factor matrix has the form

$$\mathbf{F} = \mathbf{P} \mathbf{\Lambda}^{\frac{1}{2}}. \quad (43)$$

The variance of the factor scores is equal to the eigenvalues

$$\mathbf{F}^T \mathbf{F} = \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{P}^T \mathbf{P} \mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{\Lambda}. \quad (44)$$

Taking into account that the sum of the eigenvalues is equal to the trace of $\mathbf{X}^T \mathbf{X}$, this shows that the first factor scores “extract” as much of the variance of the original data as possible, that the second factor scores

extract as much of the variance left unexplained by the first factor, and so on for the remaining factors. Incidentally, the diagonal elements of the matrix $\mathbf{\Lambda}^{\frac{1}{2}}$ that are the standard deviations of the factor scores are called the *singular values* of matrix $\mathbf{X}$.

—Hervé Abdi

See also Correspondence Analysis; Discriminant Analysis; DISTATIS; Eigenvalues; Exploratory Factor Analysis; Factor Analysis; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Multiple Factor Analysis; Multivariate Analysis of Variance (MANOVA); Partial Least Square Regression; Principal Component Analysis; Singular and Generalized Singular Value Decomposition; STATIS

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H., & Valentin, D. (2006). *Mathématiques pour les sciences cognitives* [Mathematics for cognitive sciences]. Grenoble, France: Presses Universitaires de Grenoble.
- Harris, R. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Erlbaum.
- Strang, G. (2003). *Introduction to linear algebra*. Wellesley, MA: Wellesley-Cambridge Press.

EIGENVALUES

Given a real (or complex) $p \times p$ matrix, $\mathbf{A}$, for what vectors, $\mathbf{x}_{p \times 1} \neq \mathbf{0}_{p \times 1}$, and for what scalars, λ , is it true that

$$\mathbf{Ax} = \lambda \mathbf{x} ? \quad (1)$$

A nonzero vector satisfying the above equality is called an *eigenvector* (also characteristic vector or latent root) of $\mathbf{A}$, and the associated value, λ , is called an *eigenvalue* (also characteristic root or latent root) of $\mathbf{A}$. Equation 1 holds if and only if

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}, \quad (2)$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. This equation represents a system of p equations in p unknowns, which has a unique solution if and only if

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0, \quad (3)$$

where Equation 3 is called the *characteristic equation*. Thus, a square matrix of full rank will have p -unique (i.e., different) eigenvalues λ_i , with associated eigenvectors (normalized to have unit length) $\mathbf{e}_i$. Furthermore, the set of p eigenvectors of order $p \times 1$ are all mutually orthogonal. If $\mathbf{A}$ is not of full rank, then some of the eigenvalues will be redundant and equal zero.

In the social and behavioral sciences, the eigenvalues that researchers are usually most interested in are those of the variance-covariance matrix, $\mathbf{\Sigma}$. Although, in the absence of redundant variables, it is uncommon to observe eigenvalues equal to zero, it is often the case that some of the eigenvalues of $\mathbf{\Sigma}$ will be close to zero. Ordering the eigenvalues from largest to smallest,

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p,$$

a primary interest is reducing the variable space from p dimensions to k dimensions, where k is some dimensionality that retains an adequate amount of the variability present in the original dimensionality. Furthermore, the original $n \times p$ data matrix, $\mathbf{X}$, is reduced to k -dimensional space by the operation $\mathbf{X}_{n \times p} \mathbf{e}_{p \times k}^T = \mathbf{X}_{n \times k}^*$. Thus, the k -transformed variables in $\mathbf{X}^*$ are each a linear combination of the original p variables. Finally, since all eigenvectors are orthogonal, the transformed variables in $\mathbf{X}^*$ are uncorrelated.

To understand the relevance of the data reduction process, it is important to realize the following two results:

$$\text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^p \Sigma_{ii} = \sum_{i=1}^p \lambda_i$$

and

$$\mathbf{\Sigma} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2' + \dots + \lambda_p \mathbf{e}_p \mathbf{e}_p'.$$

The first result indicates that the larger eigenvalues account for more of the variance in the p -dimensional space, while the second result indicates that the original variance-covariance matrix, $\mathbf{\Sigma}$, can be represented as a sum of p matrices all of size $p \times p$ (this decomposition is known as the *spectral decomposition*). Popular techniques such as principal components analysis and some types of factor analysis (which often

include rotation of the eigenvectors to introduce interpretability and/or correlation into the reduced dimensionality) reduce the number of dimensions from p to k such that an adequate percentage of variance,

$$\frac{\sum_{i=1}^k \lambda_i}{\text{tr}(\Sigma)},$$

is explained. Some other multivariate techniques relying on eigenvalues and eigenvectors are canonical correlation, cluster analysis (some types), correspondence analysis, multiple discriminant analysis, and multivariate analysis of variance.

—Douglas Steinley

See also Factor Analysis; Multiple Factor Analysis

Further Reading

- Carroll, J. D., & Green, P. E. (1997). *Mathematical tools for applied multivariate analysis*. San Diego, CA: Academic Press.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Leon, S. J. (1998). *Linear algebra with applications* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

EM ALGORITHM

One central goal of statistical inference is to estimate parameters of a model. In a typical statistical analysis, a likelihood function, $L(\text{data}; \theta)$, is used to describe the relationship between the data and some unknown parameters θ that control the behavior of the data. A good estimate of θ is a value that maximizes the likelihood function, for example, the “most likely” value supported by the data. This estimate is called the *maximum likelihood estimate* (MLE). In some cases, we can find the MLEs analytically. But more often, we need to use numerical methods. The EM algorithm is one of the numerical techniques for this purpose.

The *EM algorithm* is an iterative algorithm for finding the MLEs of the parameters when the data are incomplete. The term *incomplete* refers to two

situations: The first occurs when some of the data are *missing*, due to difficulties in data collection process for example. The second situation occurs when the direct optimization of the likelihood function is difficult but when *latent* parameters are added, the problem becomes more tractable.

Although the two situations might sound different in description, from a statistical point of view, they are similar and share the same features. First, a complete data set is divided into an “observed” part and an “unobserved” part. The unobserved part can be either missing or latent. Second, direct optimization of the likelihood function based on the observed data might be difficult, but it becomes manageable with the likelihood function based on the complete data. The EM algorithm provides a bridge to find the MLEs of the observed data using the complete data likelihood.

EM stands for *expectation* and *maximization*. They are the two essential steps in the algorithm: the expectation step (E-step) and the maximization step (M-step). The intuition of the EM algorithm is simple: We first guess the unobserved values using their expected values (E-step) and then pretend the guessed values were observed a priori and proceed to estimate the parameters based on the complete data (M-step). Because the M-step provides us with new estimates of the parameters, we again guess the unobserved values based on these new estimates. This iterative process is repeated until convergence; for example, two consecutive estimates of the parameters yield very close values.

This idea had been widely adapted in various disciplines for a long period of time, although the most frequent citation of the EM algorithm was made by Dempster, Laird, and Rubin in 1977. They provided a rigorous framework to implement this idea: That is, the correct procedure is not to impute the individual unobserved observations, but instead the complete data sufficient statistics, or more generally the (log) likelihood function itself, by the conditional expectation given the observed data. There had also been many similar formulations of the same idea prior to this paper.

The formulation of the EM algorithm is as follows. Let Y_{obs} and Y_{unobs} be the observed and unobserved data, respectively. The observed data likelihood is $L(Y_{\text{obs}}; \theta)$, and the complete data likelihood is

$L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta)$. The EM algorithm obtains $\hat{\theta}$ for $L(Y_{\text{obs}}; \theta)$ via the following steps:

1. Take an initial guess of the parameter values $\theta^{(t)}$, with $t = 0$.
2. E-step: Compute the conditional expected value of the log of the complete data likelihood, $\log L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta)$, given the observed data Y_{obs} , assuming $\theta = \theta^{(t)}$:

$$Q(\theta; \theta^{(t)}) = E[\log L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta) | Y_{\text{obs}}; \theta = \theta^{(t)}]. \quad (1)$$

The expectation is calculated based on the density

$$f(Y_{\text{unobs}} | Y_{\text{obs}}, \theta = \theta^{(t)}).$$

3. M-step: Determine a new value $\theta^{(t+1)}$ that maximizes $Q(\theta; \theta^{(t)})$.

Repeat Steps 1 and 2 until $\|\theta^{(t+1)} - \theta^{(t)}\| \leq \varepsilon$, where ε is a certain predetermined precision value and $\|\cdot\|$ represents a certain distance measure (for example, it can be the Euclidean distance, or maximum absolute difference between two vectors, if θ is high dimensional). Once this criterion is satisfied, the EM algorithm has converged.

The EM algorithm results in a sequence of $\theta^{(t)}$ that always increases the values of the likelihood function, $L(Y_{\text{obs}}; \theta^{(t+1)}) \geq L(Y_{\text{obs}}; \theta^{(t)})$. Verification of this is beyond the scope of this entry. References can be found in many standard textbooks.

In practice, this monotonicity property implies that an EM algorithm can converge only to a local maximum of the observed likelihood function. Therefore, it is essential to either choose a good starting value for $\theta^{(0)}$ or, if that is not possible, repeat the algorithm a number of times with starting points that are far apart. This would safeguard against the possibility of reaching only a local maximum estimate.

We illustrate the algorithm using a simple example adopted from Rao, from 1973. Suppose that the complete data have five data points, y_1, y_2, y_3, y_4, y_5 , and the data have a multivariate distribution with probabilities

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

where $0 \leq \theta \leq 1$. The complete data likelihood function is

$$\begin{aligned} L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta) &= \frac{N!}{\prod_{i=1}^5 y_i!} \left(\frac{1}{2}\right)^{y_1} \left(\frac{\theta}{4}\right)^{y_2} \\ &\quad \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{1-\theta}{4}\right)^{y_4} \left(\frac{\theta}{4}\right)^{y_5}, \end{aligned} \quad (2)$$

where

$$N = \sum_{i=1}^5 y_i.$$

This is a straightforward estimation problem if all the data points are available. We can obtain the MLE of θ using Equation 2 by first taking the derivative with respect to θ :

$$\frac{\partial L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta)}{\partial \theta} = \frac{(y_2 + y_5)}{\theta} - \frac{(y_3 + y_4)}{1 - \theta}. \quad (3)$$

Setting Equation 3 to zero, we obtain a solution for θ :

$$\frac{y_2 + y_5}{y_2 + y_3 + y_4 + y_5}. \quad (4)$$

However, suppose that instead of five data points, we observe only four counts: $Y_{\text{obs}} = (y_1 + y_2, y_3, y_4, y_5) = (125, 18, 20, 34)$. The data point y_1 (or y_2) is not observed directly. Based on these four data points, we have an observed data likelihood function $L(Y_{\text{obs}}; \theta)$:

$$\begin{aligned} L(Y_{\text{obs}}; \theta) &\propto \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1 + y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \\ &\quad \left(\frac{1-\theta}{4}\right)^{y_4} \left(\frac{\theta}{4}\right)^{y_5}. \end{aligned} \quad (5)$$

We can still obtain the MLE of θ using Equation 5 by taking the derivative

$$\frac{\partial L(Y_{\text{obs}}; \theta)}{\partial \theta} = \frac{y_1 + y_2}{2 + \theta} + \frac{y_5}{\theta} - \frac{y_3 + y_4}{1 - \theta}, \quad (6)$$

setting it to zero, and solving for θ . However, instead of working with a linear function based on Equation 3, we now have a quadratic function in θ that is more difficult to obtain an answer to:

$$(y_1 + y_2)\theta(1 - \theta) + y_5(2 + \theta)(1 - \theta) - (y_3 + y_4)\theta(2 + \theta) = 0. \quad (7)$$

Here is a situation where it is easier to optimize $L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta)$, Equation 2, than $L(Y_{\text{obs}}; \theta)$, Equation 5. And this calls for the use of the EM algorithm.

The E-step works as follows. By Equation 1, given the observed $Y_{\text{obs}} = \{y_1 + y_2 = 125, y_3 = 18, y_4 = 20, y_5 = 34\}$ and $\theta^{(t)}$, we have

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= E(\log L(Y_{\text{obs}}, Y_{\text{unobs}}; \theta) | Y_{\text{obs}}; \theta = \theta^{(t)}) \\ &= E\left\{C + y_2 \log\left(\frac{\theta}{4}\right) + (y_3 + y_4) \log\left(\frac{1 - \theta}{4}\right) \right. \\ &\quad \left. + y_5 \log\left(\frac{\theta}{4}\right) | Y_{\text{obs}}; \theta = \theta^{(t)}\right\} \\ &= C + E(y_2 | y_1 + y_2 = 125; \theta = \theta^{(t)}) \log\left(\frac{\theta}{4}\right) \\ &\quad + E(y_3 + y_4 | y_3 = 18, y_4 = 20) \log\left(\frac{1 - \theta}{4}\right) \\ &\quad + E(y_5 | y_5 = 34) \log\left(\frac{\theta}{4}\right) \\ &= C + 125 \times \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4} \log\left(\frac{\theta}{4}\right) \\ &\quad + 38 \times \log\left(\frac{1 - \theta}{4}\right) + 34 \times \log\left(\frac{\theta}{4}\right). \end{aligned} \quad (8)$$

C represents a constant term in $Q(\theta; \theta^{(t)})$ that is independent of the parameter. Note the distribution of y_2 given that $y_1 + y_2 = 125$ follows a binomial $(125, \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4})$. (That is, a binomial distribution with 125 number of trials and the “success” probability being $\frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4}$). Hence,

$$\begin{aligned} E(y_2 | y_1 + y_2 = 125; \theta = \theta^{(t)}) \\ = 125 \times \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4}. \end{aligned}$$

The M-step is also easy. Using the same technique as we used to find the MLE based on the complete data likelihood (Equations 3 and 4), the maximum of Equation 8 can be found at

$$\theta^{(t+1)} = \frac{125 \times \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4} + 34}{125 \times \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4} + 72}.$$

We initialized the EM algorithm with a starting value of $\theta^{(0)} = 0.5$ and used a stopping criterion of $|\theta^{(t+1)} - \theta^{(t)}| < 10^{-8}$. Table 1 lists iteration numbers, estimated values of $\theta^{(t)}$, and stepwise differences. The stopping criterion is satisfied after 10 iterations, with $\hat{\theta} = 0.62682150$. Different starting values of $\theta^{(0)}$ do not change the result.

Analytical solution of Equation 7 yields an MLE of θ that is very close to the EM estimate:

$$\hat{\theta} = \frac{15 + \sqrt{15^2 + 4 \times 197 \times 68}}{2 \times 197} \approx 0.6268215.$$

Currently, there is no software that implements the EM algorithm for a general problem, and almost

Table 1 Estimated Values of θ in the EM Iterations

t	$\theta^{(t)}$	$ \theta^{(t)} - \theta^{(t-1)} $
0	0.50000000	—
1	0.60824742	0.1082474
2	0.62432105	0.01607363
3	0.62648888	0.002167829
4	0.62677732	0.0002884433
5	0.62681563	3.830976e-05
6	0.62682072	5.086909e-06
7	0.62682140	6.754367e-07
8	0.62682142	8.968368e-08
9	0.62682149	1.190809e-08
10	0.62682150	1.581141e-09

every problem requires both tailored implementations and careful personal monitoring (e.g., starting values and convergence). Although some problems can be solved efficiently using high-level but relatively slow statistical languages, such as R or Splus, more complicated problems can potentially take a long time to complete, both in human effort and in computational resources. Various attempts have been proposed to improve the speed of the EM algorithm. One direction involves the direct extension of the original EM algorithm. These approaches include the expectation-conditional maximization (ECM), the expectation-conditional maximization either (ECME), the space-alternating generalized EM algorithm (SAGE), the alternating expectation-conditional maximization (AECM), and the parameter-expanded EM algorithm (PX-EM). Another school of thought for speeding up convergence is to combine EM with various numerical acceleration techniques. These approaches include combining EM with (a) Aitken's acceleration method, (b) Newton-type method, and (c) conjugate-gradient acceleration method.

Finally, the EM algorithm presented in this entry provides us with only an MLE of the parameter. There exist modifications that augment the EM algorithm with some computations to produce the standard errors of the MLEs. The standard errors of the estimates are often estimated via asymptotic theory.

—Jung-Ying Tzeng

See also Inferential Statistics

Further Reading

- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Fessler, J. A., & Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 4, 2664–2677.
- Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, 62, 257–270.
- Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5, 1–18.
- Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633–648.
- Liu, C., Rubin, D. B., & Wu, Y. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85, 755–770.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44, 226–233.
- McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions* (Wiley Series in Probability and Statistics). New York: Wiley.
- Meilijson, I. (1989). A fast improvement of the EM algorithm in its own terms. *Journal of the Royal Statistical Society, Series B*, 51, 127–138.
- Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267–278.
- Meng, X. L., & van Dyk, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59, 511–567.
- Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 355–368). Cambridge: MIT Press.
- Rao, C. R. (1973). *Linear statistical inference and applications*. New York: Wiley.
- Tanner, M. (1996). *Tools for statistical inference*. New York: Springer-Verlag.

EMBEDDED FIGURES TEST

The Embedded Figures Test (EFT) is a measure of individual differences in how surrounding fields influence a person's perception (i.e., the ability to avoid the confusion of conflicting perceptual cues). Although the EFT is a cognitive task, its relation to personality is the primary interest. During his research on perception, Witkin noticed that people varied markedly in their abilities to perform on the Rod and Frame Test and in a task judging body orientation in a tilted room. To demonstrate these same perceptual differences in a paper-and-pencil format, Witkin chose materials Gottschaldt used in his studies of the role of past experiences in perception. For these materials, one identifies a simple figure, previously seen, within a larger, more complex figure. Witkin chose 8 of Gottschaldt's simple figures and 24 of his complex figures. Preliminary experiments demonstrated an insufficient number of difficult examples using this material. Using the same principles of patterning to

create new figures proved ineffective, so Witkin used coloring to reinforce patterns. This obscured the simple patterns and increased difficulty. The primary goals in the final selection of materials for the EFT were achieving a graded difficulty and containing sufficient variety of simple figures to reduce the potential for practice effects.

The EFT is administered individually and consists of 24 trials, each using different complex figures and never using the same simple figures in 2 successive trials. During each trial, the figures are presented separately in the sequence of complex figure, simple figure, then complex figure. This pattern is used to impress upon the participant the complex figure and discourage the participant from concentrating on the simple figure at its expense, thereby increasing difficulty. The administrator notes the time at which the participant verbally indicates he or she has identified the simple figure and continues timing until the participant successfully traces it within the complex figure. The score is the time at which the participant verbally indicates he or she has identified the simple figure, provided it is confirmed correct. The total score is the summation of the time to complete all trials. Lower scores are considered field independent, and higher scores are considered field dependent.

It has been shown that people are consistent across trials in their abilities to locate simple figures, indicating that personal factors, not the structure of the field alone, are responsible for the individual differences observed. Also, Witkin noted a sex difference, with men outperforming women. There are relationships to other tests, including concept formation tests and intelligence tests, which have generated debate as to whether the EFT is a measure of cognitive ability or cognitive style. Relationships to measures of general intelligence support the ability thesis. Supporting the style thesis, comparisons with the Vigotsky Test show that field dependents are related to those who use the perceptual approach and field independents are related to those who use the conceptual approach.

—John R. Reddon and Shane M. Whipple

See also Personality Tests

Further Reading

Witkin, H. A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, 19, 1–15.

EQUIVALENCE TESTING

An *equivalence test* is a method of hypothesis testing that is a variation of the more commonly used method of significance testing. In *significance testing*, the idea is to test a null hypothesis that two means are equal. Rejecting the null hypothesis leads to the conclusion that the population means are significantly different from each other. *Equivalence testing*, on the other hand, is used to test a null hypothesis that two means are not equal. Rejection of the null hypothesis in an equivalence test leads to the conclusion that the population means are equivalent. The approach of equivalence testing differs from the more familiar hypothesis tests, such as the two-sample *t* test, where rejection of the null is used to infer that the population means are significantly different.

Equivalence testing originated in the fields of biostatistics and pharmacology, where one often wishes to show that two means are “equivalent” within a certain bound. Many researchers often incorrectly conclude that the failure to reject the null hypothesis in a standard hypothesis test (such as a *t* test) is “proof” that the null hypothesis is true and hence that the populations are “equivalent.” This erroneous inference neglects the possibility that the failure to reject the null is often merely indicative of a Type II error, particularly when the sample sizes being used are small and the power is low.

We will consider a common equivalence test known as the *two one-sided tests* procedure, or TOST. It is a variation of the standard independent-samples *t* test. With a TOST, the researcher will conclude that the two population means are equivalent if it can be shown that they differ by less than some constant τ , the equivalence bound, in both directions. This bound is often chosen to be the smallest difference between the means that is practically significant. Biostatisticians often have the choice for τ made for them by government regulation.

The null hypothesis for a TOST is $H_0: |\mu_1 - \mu_2| \geq \tau$. The alternative hypothesis is $H_1: |\mu_1 - \mu_2| < \tau$.

The first one-sided test seeks to show that the difference between the two means is less than or equal to $-\tau$. To do so, compute the test statistic

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 + \tau \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where s_p is the pooled standard deviation of the two samples. Then, compute the p value as $p_1 = P(t_1 < t_v)$, where t_v has a t -distribution with $\eta = n_1 + n_2 - 2$ degrees of freedom.

Similarly, the second one-sided test seeks to show that the difference between the two means is greater than or equal to $+\tau$. To do so, compute the test statistic

$$t_2 = \frac{\bar{x}_1 - \bar{x}_2 - \tau \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}.$$

Compute the p value as $p_2 = P(t_2 > t_v)$. Then let $p = \max(p_1, p_2)$ and reject the null hypothesis of nonequivalence if $p < \alpha$.

Establishing equivalence between two treatments or groups has applications not just in biostatistical and pharmacological settings but also in many situations in the social sciences. Many hypotheses currently tested and interpreted with standard significance testing should be approached with equivalence testing.

—Christopher J. Mecklin and Nathaniel R. Hirtz

See also Hypothesis and Hypothesis Testing; Null Hypothesis Significance Testing

Further Reading

- Berger, R., & Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11, 283–319.
- Blair, R. C., & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistics*, 1, 139–142.
- Rogers, J., Howard, K., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565.

Schirmann, D. (1981). On hypothesis testing to determine if the mean of the normal distribution is contained in a known interval. *Biometrics*, 37, 617.

Westlake, W. (1979). Statistical aspects of comparative bioequivalence trials. *Biometrics*, 35, 273–280.

ESSAY ITEMS

Essay items require the test taker to write a coherent and informative response to a question, with the purpose of assessing how well the test taker can organize information and express his or her ideas in writing.

Essay questions can be open-ended (also called unrestricted or extended) questions or closed-ended (also called restricted) questions. An *open-ended essay question* is one in which there are no restrictions on the response, including the amount of time allowed to finish, the number of pages written, or material included. A *closed-ended question* is one in which there are restrictions on a response.

Guidelines for writing an essay question are as follows:

1. Adequate time should be allowed to answer the question. By their very design, essay questions can take a considerable amount of time to answer. Regardless of whether an essay question is closed- or open-ended, the test preparer must know how much time will be allowed, as must the test taker.
2. The essay question needs to be complete and clear.
3. The same essay question should be administered to all test takers. This reduces the burden placed on the developer of the test questions in terms of time needed to create more than one item but also reduces the likelihood that questions on the same topic are of different levels of difficulty.

Advantages and Disadvantages of Using Essay Items

Essay items have several advantages. First, they are the best way of finding out what the test taker knows and also how well the test taker can relate ideas to one another. Second, security is increased, since it is

very difficult to plagiarize during an essay item examination. Finally, and this is very important, if the test constructor knows the material well, essay questions can effectively tap higher-order learning.

However, there are disadvantages to essay items as well. First, they emphasize writing and do not necessarily tap the test taker who is knowledgeable about ideas and their relationships to one another but just cannot express this in words. Second, it is difficult for essay questions to adequately sample the entire universe of what the test taker might have learned. Third, essay questions are not easy to score, with even a small number of items and a small number of test takers resulting in a large number of essays to read and grade.

Scoring Essay Items

Scorers should provide plenty of time to score an essay item. Each item has to be read and then scored, and often the scorer reads the items more than once, the first time for a general overview of the content and the second time for a more detailed analysis, including an assessment of content (again) and writing skills (such as grammar, transitions, and sentence usage).

A model of a correct answer should also be used to serve as a basis for comparison. Having a model greatly increases the likelihood that the scorer will evaluate each answer fairly and have as objective a standard as is possible, since the scorer can compare what is there (the test taker's response) to what should be there (the model response).

All items should also be scored across all test takers. The model answer for Question #1, for example, should be used to score Item #1 for all test takers. This allows the scorer not only to make absolute judgments in comparison to the model answer but also to make relative judgments (if necessary) within any one item.

Finally, responses should be graded without knowing the test taker's identity. Since there is a subjective element that can enter into the grading of essay questions, not knowing who the test taker is (and avoiding that possible bias) can be a great help.

— Neil J. Salkind

See also Multiple-Choice Items; Standards for Educational and Psychological Testing

Further Reading

Salkind, N. J. (2006). *Tests and measurement for people who (think they) hate tests and measurements*. Thousand Oaks, CA: Sage.

Computer grading of essays: http://www.salon.com/tech/feature/1999/05/25/computer_grading/

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Curren, R. R. (2004). Educational measurement and knowledge of other minds. *Theory and Research in Education*, 2(3), 235–253.

This article addresses the capacity of high-stakes tests to measure the most significant kinds of learning and discusses the value of several different test items, such as multiple choice and essay. It begins by examining a set of philosophical arguments pertaining to construct validity and alleged conceptual obstacles to attributing specific knowledge and skills to learners, and it continues to examine the difficulties involved in combining adequate validity and reliability in one test. The literature on test item formats is discussed as it relates to the potential validity of multiple-choice items, and the rater reliability of constructed-response items (such as **essay items**) is addressed through discussion of the methods used by the Educational Testing Service and a summary report of alternative methods developed by the author.

ESTIMATES OF THE POPULATION MEDIAN

The median (θ), the point on a scale below which 50% of the observations fall, is an ancient but commonly used measure of central tendency or location parameter of a population. The sample median can be written as

$$M_s = (1 - k)X_{(i)} + kX_{(i+1)}, \quad (1)$$

where $i = [(n+1)/2]$ and $k = \{(n+1)/2\}$ are the whole and decimal portions of the $(n+1)/2$, respectively.

The sample median, however, suffers from several limitations. First, its sampling distribution is intractable, which precludes straightforward development of an inferential statistic based on a sample median.

Second, the sample median lacks one of the fundamental niceties of any sample statistic. It is not the best unbiased estimate of the population median. Indeed, a potentially infinite number of sample statistics may more closely estimate the population median.

One of the most commonly used competitors of the sample median is the Harrell-Davis estimator, from 1982, which is based on Maritz and Jarrett, from 1978. Let $X = (X_1, \dots, X_n)$ be a random sample of size n and $\tilde{X} = (X_{(1)}, \dots, X_{(n)})$ be its order statistics ($X_{(1)} \leq \dots \leq X_{(n)}$). The estimator for p th population quantile takes the form of a weighted sum of order statistics with the weights based on incomplete beta function:

$$M_{HD} = \sum_{i=1}^n W_{n,i}^{HD} X_{(i)}, \quad (2)$$

where the weights $W_{n,i}^{HD}$ can be expressed as

$$W_{n,i}^{HD} = I_{i/n}((n+1)/2, (n+1)/2) - I_{(i-1)/n}((n+1)/2, (n+1)/2), \quad (3)$$

where $i = 1, \dots, n$.

An interesting property of Equation 3 is that the resulting beta deviates represent the approximation of the probability that the i th-order statistic is the value of the population median. However, that observation is irrelevant to the task of finding the best estimate of the population median (or any specific quantile). In other words, this observation neither proves that the Harrell-Davis is the best estimator nor precludes the possibility that other multipliers may be substituted for Equation 3 in Equation 2 that produce a closer estimate of the population median.

A new competitor was recently proposed by Shulkin and Sawilowsky, in 2006, which is based on a modified double-exponential distribution. Calculate the weights $W_{n,i}^{AltExp}$ in the following form:

$$W_{n,i}^{AltExp} = \int_{-\infty}^{-n/3+2i/3} \frac{1}{2} (1 + \operatorname{sgn}(x)(1 - e^{-8|x|/n})) dx - \int_{-\infty}^{-n/3+2(i-1)/3} \frac{1}{2} (1 + \operatorname{sgn}(x)(1 - e^{-8|x|/n})) dx, \quad (4)$$

where $i = 1, \dots, n$.

The weights in Equation 4 can be interpreted as the probability that a random variable falls between $-n/3 + 2(i-1)/3$ and $-n/3 + 2i/3$. The modified form of the Laplace distribution used here was obtained through a series of Monte Carlo minimization studies. The estimate is calculated as a weighted sum,

$$M_{AltExp} = \sum_{i=1}^n W_{n,i}^{AltExp} X_{(i)}. \quad (5)$$

There are two ways to judge which competitor is superior in estimating the population median regardless of distribution or sample size. One benchmark is the smallest root mean square error from the population median. Another is the closeness to the population median.

Let M^P be the population median. Let N_R be a number of Monte-Carlo repetitions and M_i^j be the median estimate by the j th method in i th repetition, $j = 1, \dots, N_M$. Here N_M is the number of methods. Then, mean square error (MSE) can be defined as follows:

$$\varepsilon_{MSE}^j = \frac{\sqrt{\sum_{i=1}^{N_R} (M^P - M_i^j)^2}}{N_R}. \quad (6)$$

Further, calculate deviation of each estimate from the population median:

$$\Delta M_i^j = |M^P - M_i^j|, j = 1, \dots, N_M, i = 1, \dots, N_R. \quad (7)$$

For each $i = 1, \dots, N_R$, find a set of indexes $I(j)$, $j = 1, \dots, N_M$, such that

$$\Delta M_i^{I(1)} \leq \Delta M_i^{I(2)} \leq \dots \leq \Delta M_i^{I(N_M)}. \quad (8)$$

The rank-based error (RBE) can now be defined as follows:

$$\varepsilon_{RBE}^j = \frac{\sum_{i=1}^{N_R} (I(j) - 1)/N_M}{N_R}. \quad (9)$$

A Monte Carlo study was conducted to compare these three sample statistics. The distributions that

were sampled included the standard normal (De Moivre or Gauss), uniform, exponential ($\mu = \sigma = 1$), chi-squared ($df = 2$), and Student's t ($df = 3$). The sample sizes were $n = 5, 10, 15, 20, 25, 30$, and 50 . Results showed that the modified double exponential minimizes the mean square error from the population median, followed by the Harrell-Davis estimator, and, finally, the sample median. The modified double exponential had the largest frequency of occurrences of being the closest to the population median, with the Harrell-Davis and the sample median obtaining fewer occurrences of being the closest, respectively.

Example

Let $X = (10, 12, 13, 15, 20)$ be a random sample of size $n = 5$, drawn from an unknown population. The task is to estimate the population median based on these data points. $\tilde{X} = (10, 12, 13, 15, 20)$ is its order statistic.

Sample Median

The sample median is $M = 13$. This result is available in most computer statistics packages. For example, in SPSS, the commands are Analyze | Descriptive Statistics | Explore.

Harrell-Davis

The weights are available by taking expected values for size n from the beta distribution. In this example, the weights are $W_{n,1}^{HD} = .0579$, $W_{n,2}^{HD} = .2595$, $W_{n,3}^{HD} = .3651$, $W_{n,4}^{HD} = .2595$, and $W_{n,5}^{HD} = .0579$. Thus,

$$M_{HD} = .0579 \times 10 + .2595 \times 12 + .3651 \times 13 + .2595 \times 15 + .0579 \times 20 = 13.4912.$$

Modified Double Exponential

The weights are available in Shulkin, from 2006. In this example, they are calculated as $W_{n,1}^{AltExp} = .0662$, $W_{n,2}^{AltExp} = .1923$, $W_{n,3}^{AltExp} = .4133$, $W_{n,4}^{AltExp} = .1923$, and $W_{n,5}^{AltExp} = .0662$. Thus,

$$M_{AltExp} = .0662 \times 10 + .1923 \times 12 + .4133 \times 13 + .1923 \times 15 + .0662 \times 20 = 12.5539.$$

—Boris Shulkin and Shlomo S. Sawilowsky

See also Measures of Central Tendency; Median

Further Reading

- Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69, 635–640.
- Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73(361), 194–196.
- Shulkin, B. (2006). *Estimating a population median with a small sample*. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

ETHICAL ISSUES IN TESTING

All professional activities of psychologists, including psychological testing, are governed by ethical standards and principles, such as the ethics code of the American Psychological Association (APA). In this entry, the discussion focuses on the ethical practice of formal testing activities as outlined in the APA ethics code.

Selection and Use of Tests

Before the first test item is administered, the evaluator makes important decisions regarding the specific tests to be employed with a particular client. When evaluators select tests, they are ethically obligated to ensure that the tests fall within their areas of competence. For example, a psychologist trained exclusively to work with children will probably be adequately trained to administer children's IQ tests but may need additional training to reach a level of competence with adult IQ tests. Also, tests should be selected for a particular evaluation only if they are appropriate for the purpose of that evaluation. Similarly, evaluators should select tests that are suitable for the client being evaluated, especially considering the client's age, cultural background, and linguistic abilities. Thus, if a psychologist's task is to conduct a personality

evaluation for which a popular test, such as the Minnesota Multiphasic Personality Inventory-Second Edition (MMPI-2), might be appropriate, the psychologist should be familiar with the age range restrictions of the various versions of the adolescent and adult forms of the test, as well as the languages in which it is available.

Evaluators should select tests that have established reliability and validity. If no such test is available and the evaluator chooses to use a test with questionable or unknown reliability and validity, this fact should be noted in the report of the results. Likewise, evaluators should use tests in accordance with the purpose and administration procedure outlined in the tests' manuals. This is particularly important with standardized face-to-face tests, such as the Wechsler IQ tests, where uniform administration and scoring are essential to the validity of the test results.

Informed Consent Regarding Testing

Also, before the first test item is administered, the evaluator is ethically obligated to obtain informed consent from the client or from his or her legal guardian, when appropriate. This obligation stands unless the testing is mandated by law or governmental regulations or in other isolated cases, as explained in the APA ethics code. Even when it is not necessary to obtain informed consent, ethical evaluators still inform clients about the testing activities they are about to undergo. In practice, there is some variability among evaluators regarding the specific information they present to a client prior to testing, but in general, this process should include an explanation of the nature and purpose of the testing, any costs or fees, the involvement of third parties (such as third-party payers, legal authorities, or employers), and the limits of confidentiality. In testing situations, confidentiality may be limited by state laws involving a psychologist's "duty to warn" or mandated child abuse reporting. Typically in these cases, a psychologist who, during testing, discovers that a client intends to cause harm to himself or herself or another individual or that a child is being abused breaks confidentiality in order to protect the individual at risk. It is also important to discuss the limits of confidentiality with minors and

their parents or guardians, especially regarding the access to testing information that the parents or guardians may have.

Clients should be informed about testing in language they can understand, and their consent should be voluntary rather than coerced. Moreover, the evaluator is ethically obligated to give the client an opportunity to ask questions and receive answers about the testing process before it begins. Generally, it is important to ensure that the client is adequately informed and agreeable to the testing process before beginning.

The Test Itself

When the creators of psychological tests design, standardize, and validate their tests, they should use appropriate psychometric procedures and up-to-date scientific knowledge. Test developers should also aim to minimize test bias as much as possible and should create a test manual that adequately educates test administrators about when, how, and with whom the test should be used.

Evaluators are ethically obligated to avoid obsolete tests. A test may become obsolete when it is replaced by a revision that represents a significant improvement in terms of psychometrics, standardization, or applicability. For example, both the child and adult versions of the Wechsler intelligence tests have been repeatedly revised, with each new edition superseding the previous edition. Likewise, the original Beck Depression Inventory was made obsolete when a revised edition was created in the 1990s to better match depression symptoms as listed in the revised *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. In other cases, a test may become obsolete without being replaced by a more current edition. Several projective tests created in the first half of the 20th century may fit this description, either because their standardization sample has become antiquated or because they no longer meet professional standards for reliable and valid tests.

Like tests themselves, the data obtained via tests can become outdated as well. For example, data collected during a child's learning disability evaluation via intelligence or achievement test remains applicable to the child for only a limited period of

time. As time passes, the child's development and education warrant that similar tests be readministered. Likewise, data obtained via neuropsychological testing may remain accurate for only a limited period of time. After this period, its optimal use may be for comparison to data collected more recently via similar tests.

Qualifications of the Evaluator

Psychological testing should be conducted only by individuals with appropriate qualifications. The evaluator must have competencies specific to the test and the client in question; merely possessing a license to practice psychology does not support unlimited use of psychological tests. An important exception to this rule is the psychological trainee under supervision. Such individuals can conduct testing for training purposes but should do so with supervision suitable to their levels of training and should inform the people they evaluate (or their parents or guardians) of their status.

Scoring and Interpretation

When scoring or interpreting psychological test results, evaluators should consider client-specific variables, such as situational, linguistic, ethnic, and cultural factors. Notes regarding interpretations made in these contexts should be included in the report.

If a psychologist utilizes a scoring or interpretation service in the process of an evaluation, the psychologist should ensure that the procedure is valid for the purpose of the particular test or evaluation. Even if the scoring or interpretation was completed by another person (or computer service), the psychologist conducting the evaluation retains professional responsibility. Those offering scoring or interpretation services to other professionals should nonetheless create reliable and valid procedures and should accurately describe their purpose, method, and applications.

Use of Test Results

Although previous editions of the APA ethics code generally prohibited the release of raw test data to

clients, the most recent edition obligates psychologists to release test data to clients (with a signed release from the client) unless substantial harm or misuse can be reasonably expected. In this context, test data include client responses to test items but not the stimuli, questions, or protocols that elicited the responses. This category also includes raw and scale scores as well as notes about the client's behavior during the testing. Without a release signed by the client, psychologists should maintain the confidentiality of test data unless required by law or court order to provide these data. It is important for those conducting testing to be familiar with state laws governing these issues, as well as relevant ethical standards.

In most cases, clients will not seek their own test data. Nonetheless, all clients are entitled to receive feedback regarding their test results. In general, ethical evaluators provide an intelligible explanation to clients (or their parents or guardians) regarding their test results, the meaning of these results, and their possible implications or consequences. In some circumstances (such as some forensic evaluations or organizational assessments), this feedback or explanation procedure may be precluded; in these cases, the evaluator should inform the client during the informed consent procedure that no explanation of results will be forthcoming.

Test Security

Psychologists and others who administer psychological tests are ethically bound to maintain the security of these tests. The APA ethics code requires that reasonable efforts should be taken to maintain the integrity and security of test materials. Individual test takers should not be able to access and review psychological tests before the test administration. When individuals have prior access to tests, test questions, or test answers, the psychometric integrity of the tests is compromised. For example, if a person were to have access to the questions contained in an IQ test beforehand, the individual's test scores could be artificially inflated. Such prior access to test materials would make the test administration invalid. This breach in test security could lead to a gradual weakening in the

validity of the test in question if the test stimuli were shared with other potential test takers.

Professionals who are responsible for psychological tests should take reasonable steps to make sure that individuals are not able to review tests before administration, keep scoring keys and test materials secure, and not allow unqualified individuals access to test materials. Copyright law should also be considered before test materials are published or disclosed. Before any portion of a copyrighted test is reproduced, permission should be gained from the publisher or copyright holder.

The security of test materials may be compromised by publishing test materials in scholarly writing, including test materials in court records, maintaining poor control of test materials in academic settings, and the unauthorized distribution or publications of the test materials through Web sites and other means. Reproducing test materials in scholarly writing could compromise test security if test items or stimuli were included in the publication. Caution should be exercised in such cases to maintain test security and adhere to copyright laws. Controlling the security of tests in court settings may be obtained by asking the court to restrict the release of subpoenaed test materials to a psychologist or other individual bound by the applicable ethical standards. Tests can be kept secure in academic settings by keeping them in a secure area and by allowing only those individuals who have been deemed competent test users to have access to the tests. However, even highly trained individuals may at times be unaware of the guidelines promulgated by test publishers that identify the different levels of training necessary for competent test use. For example, some social science researchers may use psychological tests in research that were designed for use primarily in clinical settings. However, these researchers may be unaware of the ethical guidelines that control the security of these tests. Tests designed for clinical purposes that are used in research should be maintained at a high level of security.

The Internet provides an easy method for the unauthorized distribution of test materials by individuals who are not competent test users. Furthermore, nonprofessionals are not bound by the same ethical

standards as psychologists and other test users. The unauthorized distribution or publication of test materials may not be under the control of test administrators, but test users are responsible for taking steps to avoid any opportunity for test materials and test scores to be obtained by fraudulent means.

—Andrew M. Pomerantz and Bryce F. Sullivan

See also Educational Testing Service; Ethical Principles in the Conduct of Research With Human Participants; Standards for Educational and Psychological Testing

Further Reading

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060–1073.
- Fisher, C. B. (2003). *Decoding the ethics code: A practical guide for psychologists*. Thousand Oaks, CA: Sage.
- Koocher, G. P., & Rey-Casserly, C. M. (2002). Ethical issues in psychological assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of assessment psychology*. New York: Wiley.
- APA Ethics Code, including a section on assessment: <http://www.apa.org/ethics/code2002.html>
- APA statement on test security in educational settings: <http://www.apa.org/science/securetests.html>
- Test security: <http://www.apa.org/journals/amp/testsecurity.html>

ETHICAL PRINCIPLES IN THE CONDUCT OF RESEARCH WITH HUMAN PARTICIPANTS

Ethics is the study of assumptions believed to assist in distinguishing between right and wrong for the purpose of making sound moral judgments. *Ethical principles* are standards or rules that may serve as policy for determining modes of action in situations that involve or require moral judgment and decision

making. The conduct of scientific research using human participants necessarily involves ethical decision making and is rife with potential for ethical conflict. Largely in response to flagrant occurrences of unethical research with human participants, professional organizations and government agencies began specifying ethical principles to guide researchers in the mid-20th century. These principles vary somewhat but typically emphasize beneficence and nonmaleficence; fidelity, responsibility, and trust; integrity; justice; and respect for the dignity and autonomy of persons. The complexities of many ethical decisions require more than the rigid application of rules; researchers are responsible for using sound, well-reasoned judgment in planning and implementing research in a way that maximizes benefits, minimizes harm, and promotes the dignity and worth of all human participants.

History

Although ethical considerations have influenced researchers on an informal and individual basis throughout history, the first formal recognition of the importance of ethical principles in research occurred in 1947, after the Nuremberg Trials of Nazi war criminals. These trials revealed to the public that during World War II, physicians and scientists had conducted biomedical experiments on involuntary participants drawn from Nazi concentration camps. Some of these experiments were designed to assess human responses to poisons, extreme temperatures, and infections, and they essentially resembled torture. Noting that there was at the time no international law or ethics code to refer to in addressing such egregious treatment of human participants, Leo Alexander, an American physician and consultant to the prosecution during the trials, submitted a report that presented standards for legitimate, ethical research. This report formed the basis of the subsequently developed Nuremberg Code, the first formal code of ethical principles addressing the conduct of research with human participants. The Nuremberg Code emphasized principles such as informed consent, avoidance of harm, the necessity of researchers having appropriate training, and freedom of participants to withdraw at any time.

The Nuremberg Code played a significant role in shaping the content of ethical guidelines published by professional organizations such as the American Psychological Association (APA). The APA first published a general ethics code for psychologists in 1953. In 1966, APA established an ad hoc committee to further examine research ethics. In 1973, the committee published a booklet titled “Ethical Principles in the Conduct of Research With Human Participants.” This booklet, along with APA’s general ethical guidelines, has subsequently undergone revision. The most recent APA ethics code, which includes guidelines for research, was published in 2002.

The Nuremberg Code also influenced federal regulations that were set forth by the U.S. Congress in the National Research Act of 1974. This legislation created a National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research and required the formation of an institutional, or internal, review board (IRB) by every university or other organization that receives federal funds for research. The purpose of the IRB is to review proposals for research with the aim of preventing ethically questionable studies from being conducted with human participants. During the mid-1970s, the National Commission held hearings on a series of ethically problematic research efforts, including the Tuskegee Syphilis Study, which examined the degenerative course of syphilis in rural, underprivileged African American males by deliberately withholding treatment of the disease. These hearings led the National Commission to develop specific recommendations for research with human participants, published in “The Belmont Report,” which set the framework for federal regulation of research. The U.S. Department of Health and Human Services issued a set of regulations in 1981, called “Protection of Human Subjects.” This document and its subsequent revisions have continued to emphasize the principles of beneficence, justice, and respect for persons, as outlined in the Belmont Report.

Philosophical Approaches to Ethical Decisions

The ethical guidelines currently in use by biomedical and behavioral research communities have the

Western philosophical underpinnings of the deontological and utilitarian traditions. The deontological tradition, based on the assumption that ethics reflects a universal moral code, emphasizes respect for the autonomy of the individual. The utilitarian tradition, based on the assumption that moral judgments depend on the consequences of particular actions, emphasizes an optimal balance of the possible harms and potential benefits to people. The shared goal embedded in these approaches is to uphold the welfare and protection of individuals and groups by respecting the intrinsic worth and dignity of all persons and by carefully weighing the pros and cons of potential courses of action.

Cost-Benefit Considerations

Ethics codes such as those adopted by the APA generally emphasize utilitarian considerations guiding ethical decision making with respect to research. That is, instead of specifying a concrete set of rules, the guidelines require that researchers engage in a cost-benefit analysis before conducting a particular study. There are a number of possible costs of conducting research with human participants. Costs to participants may include their time, resources, and effort; possible injury, stress, anxiety, pain, social discomfort, and threats to self-esteem; and the risk of breached confidentiality. Other costs of research include the use of resources and monetary expenses required for salaries, equipment, and supplies, and possible detriment to the profession or to society, as in the case of an experimental treatment that unintentionally harms participants and causes distrust toward behavioral research.

These costs must be weighed against the possible benefits of the research. Such benefits include the advancement of basic scientific knowledge; the improvement of research or assessment methods; benefits for society, such as improved psychotherapy techniques or enhanced classroom teaching and learning processes; benefits for researchers and research trainees, such as increased knowledge and advancement toward professional and educational goals; and benefits for research participants, such as when a study testing an experimental treatment for depression

helps participants become less depressed. Researchers bear responsibility to society, science, students and trainees, and research participants. The main focus of the current ethical guidelines for research is on responsibility to participants, but this responsibility must always be held in balance with researchers' other three responsibilities.

Five Ethical Principles for Research With Human Participants

There are five general principles in the 2002 APA ethics code designed to "guide and inspire psychologists toward the very highest ethical ideals of the profession." These principles include beneficence and nonmaleficence (i.e., benefit people and do no harm); fidelity and responsibility; and integrity, justice, and respect for people's rights and dignity. The Belmont Report identified three basic ethical principles when conducting research: respect for persons, justice, and beneficence. The following are five basic ethical principles presented in the order of the general principles in the APA code that apply specifically to conducting biomedical and behavioral research with human participants.

Principle 1: Beneficence and Nonmaleficence

Representing the utilitarian tradition, this principle requires that researchers, using considerations such as those described above, strive to maximize potential benefits while minimizing risks of their research. Although the cost-benefit mandate seems straightforward, it is rarely unambiguous in practice because costs to participants and benefits to the profession and to society are difficult to accurately estimate in advance and no universally agreed-upon method or criteria exist for optimally balancing the two. Where questions arise related to the degree of risk, researchers are responsible for seeking ethical advice and implementing safeguards to protect participants. Risks that are identified in advance must be communicated to prospective research participants or their legal equivalent, and informed consent must be obtained (except in special cases approved by the

IRB, such as research involving a placebo control, in which fully informed consent compromises a scientifically required research design). Sometimes research presents risks to groups of people or social institutions. No consensus exists for whether a representative can provide consent on behalf of a collective entity, but full compliance to Principle 1 requires sensitivity to this issue.

Principle 2: Fidelity, Responsibility, and Trust

This principle requires researchers to establish and maintain a relationship of trust with research participants. For example, before individuals agree to participate in research, investigators must be clear and explicit in describing to prospective participants what they will experience and what consequences may result from participation. Researchers also are obligated to honor all promises and commitments that are made as part of the agreement to participate. When full disclosure is not made prior to obtaining informed consent (e.g., information germane to the purpose of the study would compromise its validity), safeguards must be implemented to protect the welfare and dignity of participants. In general, procedures that involve concealment or deception in a research design can be implemented only after rigorous criteria for the necessity of such procedures are met and the study is approved by the IRB. (Such instances also require a thorough debriefing of participants at the conclusion of their participation.) When children or adults with limited understanding serve as participants, researchers must implement special protective safeguards. When unintended negative consequences of research participation occur, researchers are obligated to detect, remove, and/or correct these consequences and ensure that they do not persist over time. Understandably, past ethical breaches have resulted in what some describe as widespread mistrust of biomedical and behavioral research in contemporary society. Principle 2 requires researchers to make every effort to foster trust and avoid causing further public mistrust.

Principle 3: Integrity

This principle requires researchers to “do good science,” to truthfully report their results, to take reasonable steps to correct errors that are discovered, to present work that is their own (or to otherwise make appropriate citations), to take responsibility and credit only for work that is their own, to avoid “piecemeal publication” (i.e., submitting redundant analyses of a single data set for multiple publications), to share data on which results are published with other qualified professionals provided they seek only to verify substantive claims and do not use the data for other any other purpose, and to respect the proprietary rights of others engaged in the scientific enterprise.

Principle 4: Justice

In following this principle, researchers strive for two forms of justice. The first, *distributive justice*, requires psychologists to entitle all persons equal access to the benefits of research, as well as to ensure that the risks for harm from research are not disproportionately greater for a particular group or category of persons within society. The second, *procedural justice*, refers to the adequacy of research procedures to ensure fairness, such as when easily accessible mechanisms are made available to participants to address any concerns they may have related to their participation in research.

Researchers also are promoting Principle 3 when they attend to the special concerns of underrepresented groups in developing programs of research, so as to avoid continued underinclusion and lack of representation in the knowledge base.

Principle 5: Respect for the Dignity and Autonomy of Persons

Representing the deontological tradition, this principle asserts that researchers respect research participants as human beings with intrinsic worth, whose participation is a result of their autonomous choices. The implications of this principle are far-reaching and relate to matters of obtaining informed consent,

avoiding coercive and deceptive practices, upholding confidentiality and privacy, and preserving the self-determination of participants. In abiding by this principle, psychologists are also aware of and respect individual differences, including those influenced by gender, age, culture, role, race, ethnicity, sexual orientation, religious identity, disability, linguistic background, economic status, or any other characteristic related to group membership.

Ethical Conflicts and Decision Making

The potential for ethical conflict is ubiquitous in biomedical and behavioral research. When making ethical decisions about research, it may be prudent to develop a systematic approach to reviewing all relevant sources of ethical responsibility, including one's own moral principles and personal values; cultural factors; professional ethics codes, such as the APA code; agency or employer policies; federal and state rules and regulations; and even case law or legal precedent. A process-oriented approach to ethical decision making may involve some variation of the following: (1) writing a description of the ethically relevant parameters of the situation; (2) defining the apparent dilemma; (3) progressing through the relevant sources of ethical responsibility; (4) generating alternative courses of action; (5) enumerating potential benefits and consequences of each alternative; (6) consulting with the IRB, relevant colleagues, and/or legal professionals; (7) documenting the previous six steps in the process; and (8) evaluating and taking responsibility for the results of the course of action selected. As previously mentioned, all research studies must be approved by the relevant IRB. However, approval of a research proposal by an IRB does not remove the mandate of ethical responsibility from the researcher. In making ethical decisions, researchers should consider the likelihood of self-serving bias that can lead to overestimation of the scientific value of a proposed study and underestimation of its risks.

Conclusion

Scientific research with human participants is an inherently ethical enterprise, and ethical conflicts in research are virtually inevitable. Researchers who exercise the privilege to conduct research with human participants bear the responsibility of being familiar with and abiding by the ethical principles and relevant rules and regulations established by their professional organizations and by federal and state governments. However, rigid application of rules is not a substitute for well-reasoned, responsible ethical decision making.

—Bryan J. Dik

See also Ethical Issues in Testing; Standards for Educational and Psychological Testing

Further Reading

- American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060–1073.
- Bersoff, D. N. (Ed.). (2003). *Ethical conflicts in psychology* (3rd ed.). Washington, DC: American Psychological Association.
- Miller, C. (2003). Ethical guidelines in research. In J. C. Thomas & M. Herson (Eds.), *Understanding research in clinical and counseling psychology* (pp. 271–293). Mahwah, NJ: Erlbaum.
- Office for Protection from Research Risks, Protection of Human Subjects. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research* (GPO 887-809). Washington, DC: U.S. Government Printing Office.
- Sales, B. D., & Folkman, S. (Eds.). (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.
- Sieber, J. E. (2004). Empirical research on research ethics. *Ethics and Behavior*, 14, 397–412.
- 2002 APA Ethics Code: <http://www.apa.org/ethics>
- The Belmont Report: <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm>
- Federal regulations on the protection of human participants: <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm>

EVIDENCE-BASED PRACTICE

The quest to determine what works in psychotherapy is a critical one. Evidence for therapeutic interventions can be defined in many ways. Building consensus on the definition of evidence and ensuring that *evidence-based practice* (EBP) in psychology recognizes not only the research but also the clinician's expertise and the patient's preferences, values, and culture is important to providing quality patient care and to the future of the profession of psychology. Some psychologists believe that psychological interventions should be based solely on randomized clinical trials, while others claim that other forms of evidence have their value. Regardless of their positions, most psychologists recognize that the EBP movement in U.S. society is a juggernaut, racing to achieve accountability in medicine, psychology, education, public policy, and even architecture. The zeitgeist is to require professionals to base their practice on evidence to whatever extent possible.

The American Psychological Association (APA) developed and adopted a policy statement and received a longer report on EBP in psychology at the meeting of its Council of Representatives in August 2005. The policy statement was based on the three components of the Institute of Medicine definition of EBP in medicine. Thus, the APA statement on EBP in psychology aimed to affirm the importance of attending to multiple sources of research evidence and to assert that psychological practice based on evidence is also based on clinical expertise and patient values. The statement begins, "Evidence-Based Practice in Psychology . . . is the integration of the best available research with clinical expertise in the context of patient characteristics, culture and preferences."

1. The APA policy statement has a broad view of research evidence, including multiple research designs, research in public health, health services research, and health care economics, while recognizing that there is a progression of evidence.
2. The APA policy statement explicates the competencies that make up clinical expertise. It also defines

the appropriate role of clinical expertise in treatment decision making, including attention to both the multiple streams of evidence that must be integrated by clinicians and to the heuristics and biases that can affect clinical judgment.

3. The APA policy statement articulated the role of patient values in treatment decision making, including the consideration of the role of ethnicity, race, culture, language, gender, sexual orientation, religion, age, and disability status and the issues of treatment acceptability and consumer choice.

The statement concludes,

Clinical decisions should be made in collaboration with the patient, based on the best clinically relevant evidence and with consideration of the probable costs, benefits, and available resources and options. It is the treating psychologist who makes the ultimate judgment regarding a particular intervention or treatment plan.

—Ronald F. Levant

See also Ethical Principles in the Conduct of Research With Human Participants

Further Reading

- Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.
- Norcross, J. C., Beutler, L. E., & Levant, R. F. (Eds.). (2005). *Evidence based practice in mental health: Debate and dialogue on the fundamental questions*. Washington, DC: American Psychological Association.

APA policy statement: <http://www.apa.org/practice/ebpstatement.pdf>

APA report: <http://www.apa.org/practice/ebpreport.pdf>

EXCEL SPREADSHEET FUNCTIONS

A spreadsheet function is a predefined formula. Excel, the most popular spreadsheet, has several categories of functions, including one labeled *statistical*.

One of the most simple of these functions is AVERAGE, which computes the average of a set of

Table 1 Excel Functions That Perform Statistical Operations

<i>The Function Name</i>	<i>What It Does</i>
AVERAGE	Returns the average of its arguments
CHIDIST	Returns the one-tailed probability of the chi-squared distribution
CHITEST	Returns the test for independence
CORREL	Returns the correlation coefficient between two data sets
FDIST	Returns the <i>F</i> probability distribution
FORECAST	Returns a value along a linear trend
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an <i>F</i> test
GEOMEAN	Returns the geometric mean
KURT	Returns the kurtosis of a data set
LINEST	Returns the parameters of a linear trend
MEDIAN	Returns the median of the given numbers
MODE	Returns the most common value in a data set
NORMDIST	Returns the normal cumulative distribution
NORMSDIST	Returns the standard normal cumulative distribution
PEARSON	Returns the Pearson product moment correlation coefficient
QUARTILE	Returns the quartile of a data set
SKEW	Returns the skewness of a distribution
SLOPE	Returns the slope of the linear regression line
STANDARDIZE	Returns a normalized value
STDEV	Estimates standard deviation based on a sample
STDEVA	Estimates standard deviation based on a sample, including numbers, text, and logical values
STDEVP	Calculates standard deviation based on the entire population
STDEVPA	Calculates standard deviation based on the entire population, including numbers, text, and logical values
STEYX	Returns the standard error of the predicted <i>y</i> -value for each <i>x</i> in the regression
TDIST	Returns the student's <i>t</i> distribution
TREND	Returns values along a linear trend
TTEST	Returns the probability associated with a student's <i>t</i> test
VAR	Estimates variance based on a sample
VARA	Estimates variance based on a sample, including numbers, text, and logical values
VARP	Calculates variance based on the entire population
VARPA	Calculates variance based on the entire population, including numbers, text, and logical values

values. For example, the following statement averages the numbers in cells A1 through A3:

= AVERAGE(A1:A3)

The name of the function is AVERAGE, and the argument is A1:A3n.

A similar common function produces the sum of a set of cells as follows:

= SUM(A1:A3)

In both cases, the results of these calculations are placed in the cell that contains the statement of the

	A11		fx =SUM(A1:A10)		
	A	B	C	D	
1	5				
2	6				
3	5				
4	6				
5	7				
6	8				
7	7				
8	6				
9	5				
10	6				
11	61				
12					

Figure 1 Using the SUM Function as an Example

function. For example, to use the SUM (or any other) function, follow these steps:

1. Enter the function in the cell where you want the results to appear.
2. Enter the range of cells you want the function to operate on.
3. Press the Enter key, and there you have it. Figure 1 shows the function, the argument, and the result.

Functions can be entered directly when the name of the function and its syntax are known or using the Insert command. Some selected Excel functions that perform statistical operations are shown in Table 1.

—Neil J. Salkind

See also Spreadsheet Functions

Further Reading

Instruction on using spreadsheet functions: <http://spreadsheets.about.com/od/excelfunctions/>

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) looks at data to see what they seem to say. The distribution of the observed data is examined without imposing an arbitrary probability model on it. We look for *trends*, such as patterns and linear or nonlinear relationships

between variables, and *deviations* from the trends, such as local anomalies, outliers, or clusters. This facilitates discovering the unexpected as well as confirming suspicions, rather like detective work.

EDA is sometimes viewed as a grab bag of tools, but this is a misconception. It is more accurate to view EDA as a procedure for data analysis. We start from a set of expectations or specific questions arising from the data context and explore the data with these in mind, while remaining open to observing unexpected patterns. The approach involves making many plots and numerical models of the data. Plots allow us to examine the distribution of the data without an imposed probability model; thus, statistical graphics form the backbone of EDA. Plots provide simple, digestible summaries of complex information that enable discovering unexpected structure. With the assistance of techniques such as bootstrapping, permutation, and model selection methods, we can assess whether the observed patterns in the data are more than random noise.

EDA is different from confirmatory statistical analysis. In *confirmatory analysis*, we start from a hypothesis and work to confirm or reject the hypothesis. EDA is a hypothesis discovery process. EDA provides approximate answers to any question of interest, instead of an exact answer to the wrong question. In the process of exploration, the data may suggest hypotheses, leading to follow-up confirmatory analysis with new data.

Methods in common usage that have arisen from EDA include the boxplot, stem-and-leaf plot, median polish, and projection pursuit.

History

The term *exploratory data analysis* was coined by John W. Tukey, and it is the title of his landmark book, published in 1977. It is a very idiosyncratic book, jam-packed with ways to make calculations on and draw pictures of data with paper and pencil. It is full of opinions, such as the following:

Pictures based on the exploration of data should *force* their messages upon us. Pictures that emphasize what we already know—"security blankets" to

reassure us—are frequently not worth the space they take. Pictures that have to be gone over with a reading glass to see the main point are wasteful of time and inadequate of effect. The greatest value of a picture is when it *forces* us to notice what we never expected to see.

Such opinions communicate a wisdom learned from experiences with data. The intensity of this written work, emphasized by bold and italic typeface, communicates practical advice on working with data. This integral component of Tukey's conceptualization of EDA is unfortunately missing from later treatments of EDA, which tend to make EDA look like a loose collection of ad hoc methods. In 2001, Salsburg published an easy-reading biography of Tukey's contributions on EDA in the context of other major statistical developments of the previous century.

Tukey places credit for the EDA ideas with Charles P. Winsor, who taught him “many things about data analysis that weren't in the books.” Practical issues of data analysis traverse the history of statistics. Data analysts evince the value of exploring data to see what they seem to say. EDA both descends from these matters and transcends them. IDA is the necessary prior data inspection to check that the assumptions required for formal analysis are satisfied by the data. Checking the data quality early in the analysis may save some red faces later. EDA is about the data for its own sake. There may be no need or desire for further analysis. Tukey's book may drown the reader in the author's thrill of pushing methodology in intricate dimensions, but ultimately, EDA is about the data. The methods are the tools used to dissect the data, which EDA borrows from and lends to IDA.

The Evolution of EDA

Tukey produced all of his scratched-down numbers and pictures using pencil and paper, but today, as he predicted, the computer is invaluable for EDA. EDA has evolved from a pencil-and-paper activity on tiny data sets to highly sophisticated algorithms and interactive computer-generated graphics on any size of data and any complexity of data. *Software*, a word coined by Tukey, such as R enables rapid data

calculations and making pictures easy. Software such as GGobi and Mondrian supports the use of interactive and dynamic graphics for exploring data, and emerging packages for R, such as iPlots, iSPlot, and RGtk, are enabling the integration of dynamic graphics with numerical analyses. We might call this computer-dependent approach “the new EDA.” It remains heavily dependent on statistical graphics.

EDA has expanded in many directions. For data with a spatial context, EDA has matured into exploratory spatial data analysis (ESDA). New diagnostic statistical quantities, such as local indicators for spatial dependence, and graphics, such as variocloud plots, help to find observations that are unusual in the context of their spatial neighbors. Robust methods for downweighting contamination have evolved from median polish to the expansive field of robust statistics. Quite ironically, robust approaches are sometimes described as relieving the data analyst from the task of inspecting the data. For large electronically stored databases, algorithms have emerged to mine the data for information. Examples of these algorithms are trees, forests, neural networks, and support vector machines. A statistical treatment of the area can be found in the publication by Hastie, Tibshirani, and Friedman, in 2001, including the application of bootstrapping methods to evaluate uncertainty.

EDA is permeating through statistics education. For example, introductory statistics courses, such as can be found in the publication by DeVeaux, Velleman, and Bock, in 2005, have been reframed to present statistics from an EDA perspective before introducing confirmatory methods.

An Example

This is an example derived from a case study, in 1995, by Bryant and Smith. The data on tips are collected by one waiter over a 2.5-month period at one restaurant. He recorded the tip, total bill, sex of the bill payer, smoking or nonsmoking section, size of the dining party, and the time of day. The question of interest is “What factors affect the tips?”

A basic analysis fits a multiple regression model using tip rate as the response variable to the remaining variables. It yields a model for the data with only one

significant explanatory variable, $\text{tipRate} = 0.18 - 0.01 \times \text{size}$, which can be interpreted as follows: For each increase of one person in the size of the dining party, tip rate decreases by 1%, starting from a rate of 17% for a party size of one. The model explains very little of the variation in tip rate, as you can see from the plot of the data and the model in Figure 1. The EDA approach is different from this: First, make many plots of the data and then model them. It is surprising what these deceptively casual data reveal!

To examine the variable tips, the conventional plot to use is a histogram. Tukey might have scratched up a stem-and-leaf of tips, but using the computer, a histogram is simple to produce today. EDA would suggest that several histograms of tips are generated using a variety of bin widths. Because the units are dollars and cents, a commonsense scale utilizes these units. The histograms using a full-dollar bin width show a skewed distribution, with tips centered around \$2 and few tips larger than \$6. When a bin width of 10¢ is used, it reveals something unpredicted: There are peaks at full- and half-dollar amounts. This is interesting! Are people rounding their tips? Additional observations from this plot are as follows: There are three outlying tips larger than \$7 and, surprisingly, no tips smaller than \$1.

In examining the two variables together, tips and total bill, a linear association between the two would be expected. Tips are conventionally calculated as a percentage of the bill. Total bill should explain the amount of tip. This plot is shown in Figure 1 at lower left. A linear relationship can be seen between the two variables. It is not as strong as might be expected, and there is a surprising pattern: If the plot is divided on a diagonal running from low bill/low tip to high total/high tip, there are more points in the lower right triangle. This region corresponds to tips that are lower than expected. There are very few points in the upper left triangle, where tips are higher than expected. This is also interesting! It suggests a tendency toward cheap rather than generous tips.

The data shows more unanticipated patterns when subset by the categorical variables “sex” and “smoking party.” The plots in Figure 1 (bottom right) show tip and

total bill conditioned by sex and smoking party. There is a big difference in the relationship between tip and bill in the different subsets. There is more variation in tips when the dining party is in the smoking section. The linear association is much stronger for nonsmoking parties. In the plot of female nonsmokers, with the exception of three points, the association between tip and bill is nearly perfect. The few large bills are paid mostly by males, or when paid by a female, the tips are lower than expected. The largest relative tip was paid by a male nonsmoker. These are interesting observations!

What have the data revealed? This is a small data set, but it is rich on information. It is a bit shocking and pleasing to discover so many intricate details in the numbers. Here is a summary of the observations that were made:

- Many tips are rounded to the nearest-dollar and half-dollar value.
- There are no tips less than \$1 reported.
- Tip and total bill are not as strongly associated as might be expected.
- There is more tendency toward cheap tips than generous tips.
- Smoking parties have more variation in tip and total bill.
- Males pay most of the largest bills, but when a female pays a large bill, the tip tends to be disproportionately low.
- Finally, the only factor in the data that affects tips is the size of the dining party. Tip rate has a weak negative dependence on the size of the party: The larger the party, the lower the tip rate, $\text{tipRate} = 0.18 - 0.01 \times \text{size}$.

What was accomplished in this example? A problem was posed for the collected data but was not constrained to it; many simple plots were used; and some calculations were made. The observations can be brought forward as hypotheses to be tested in confirmatory studies about tipping behavior. This is the process of exploring data.

—Dianne Cook

See also Data Mining; Graphical Statistical Methods

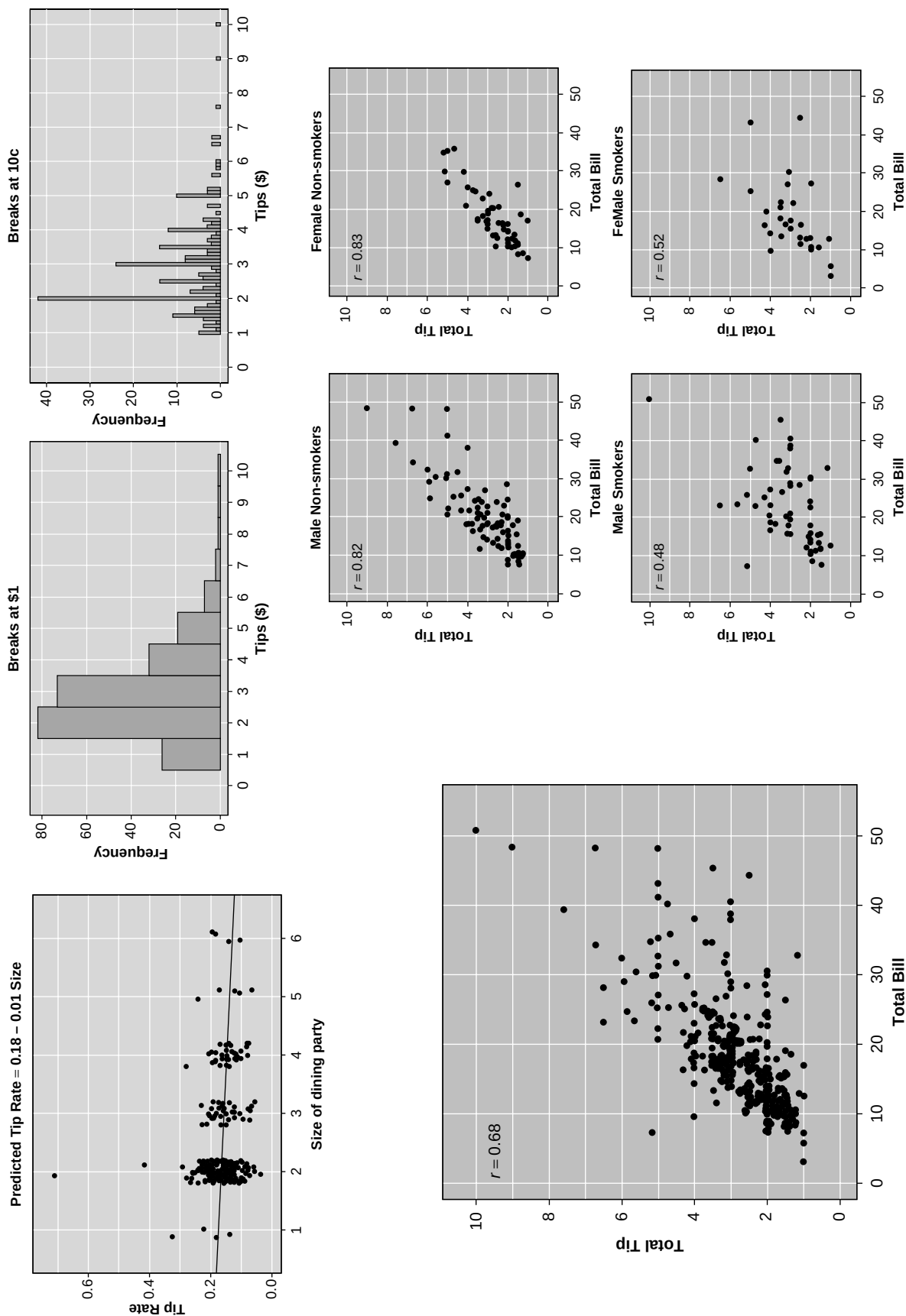


Figure 1 Plots Exploring the Tips Data

- Tip rate vs. size (jittered), to support the linear model.
- Histograms using different bin widths reveal different features of tips.
- Scatterplot showing the relationship between tip and total bill.
- The relationship between tip and total bill is much weaker for smoking parties than nonsmoking parties.

Note: These plots have been enhanced for presentation purposes. The discoveries made on the data occurred with much rougher, quickly generated graphics.

Further Reading

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Bryant, P. G., & Smith, M. A. (1995). *Practical data analysis: Case studies in business statistics*. Homewood, IL: Richard D. Irwin.
- Crowder, M. J., & Hand, D. J. (1990). *Analysis of repeated measures*. London: Chapman & Hall.
- DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2005). *Intro stats* (2nd ed.). Boston: Addison-Wesley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- R Development Core Team. (2003). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman.
- Tukey, J. (1977). *Exploratory data analysis*. Boston: Addison-Wesley.
- Vapnik, V. (1999). *The nature of statistical theory*. New York: Springer-Verlag.

EXPLORATORY FACTOR ANALYSIS

Exploratory factor analysis (EFA) is a statistical method used to explore the underlying structure of correlations among observed variables. The goal of EFA is to describe this underlying structure in a parsimonious manner by specifying a small number of factors that can account for the correlations among a set of measured variables. EFA (as opposed to *confirmatory factor analysis*) is undertaken when the researcher has no strong a priori theories about the number and nature of the underlying factors.

The mathematical basis of EFA is the common factor model, which proposes that measured variables can be explained by *underlying factors* (also called latent variables) that cannot be directly measured, but influence the measured variables. There are two types of underlying factors. *Common factors* are those that influence more than one measured variable in a set, whereas *unique factors* are those that influence only one measured variable.

The common factor model is often expressed using the following equation:

$$P = \lambda\Phi\lambda^T + D_\psi,$$

where P is the correlation matrix in the population, λ are the factor loadings (i.e., numerical values representing the strength and direction of influence of the common factors on the measured variables), Φ is the matrix of correlations among the common factors, and D_ψ is the matrix of unique factor variances (i.e., the proportion of variance in each measured variable that is explained by its unique factor).

Conducting an EFA essentially involves computing estimates for the elements in the above equation. Statistical software provides the results of such calculations, including a factor-loading matrix (λ) and a common factor correlation matrix (Φ). Programs customarily do not directly report the unique variances (D_ψ). Instead, they report the communalities (i.e., proportions of variance accounted for by the common factors), which are inversely related to the unique variances.

To illustrate the use of EFA, imagine that four ability tests (paragraph comprehension, vocabulary, arithmetic skills, and mathematical word problems) are administered. An EFA is then conducted on the correlations between these four variables. To identify and describe the underlying factors, researchers first examine the pattern of factor loadings. These values represent the magnitude and direction of influence of the common factors on the measured variables. As shown in Table 1, Paragraph Comprehension, Vocabulary, and Math Word Problems appear to load highly on Factor 1 (i.e., are strongly influenced by Factor 1), and

Table 1 Sample Factor Loading Matrix

	Factor 1	Factor 2
Paragraph Comprehension	.70	.10
Vocabulary	.70	.00
Arithmetic	.10	.70
Math Word Problems	.60	.60

Arithmetic Skills and Math Word Problems load highly on Factor 2 (i.e., are strongly influenced by Factor 2). Based on this pattern, Factor 1 might be interpreted as a verbal ability factor and Factor 2 as mathematical ability factor.

The common factor correlation matrix in Table 2 demonstrates that the verbal and mathematical ability factors are moderately correlated ($r = .41$), indicating that these factors are distinct but related constructs. The commonalities in Table 3 show that a moderate-to-large proportion of the variance in each measured variable is explained by the two common factors.

Table 2 Sample Phi Matrix

	Factor 1	Factor 2
Factor 1	—	.41
Factor 2	.70	—

Table 3 Sample Commonalities

Variable	Commonality
Paragraph Comprehension	.84
Vocabulary	.78
Arithmetic Skills	.42
Math Word Problems	.38

Decisions in Conducting EFA

Performing an EFA is a complex process that requires a researcher to make a number of decisions. For each step, researchers must choose from a variety of procedures.

The Number of Common Factors

The first decision that must be made in EFA is the appropriate number of common factors. Several statistical procedures exist to accomplish this task. These procedures are often used in combination with other considerations such as the interpretability and replicability of the factor analysis solutions.

The Kaiser Criterion

This commonly used procedure involves generating eigenvalues from the correlation matrix. *Eigenvalues* are numerical values that can be calculated from a correlation matrix and represent the variance in the measured variables accounted for by each common factor. The number of eigenvalues computed is equal to the number of measured variables. If a factor has a low eigenvalue, it does not account for much variance and can presumably be disregarded. The Kaiser criterion (also called the “eigenvalues-greater-than-1 rule”) proposes that a researcher should retain as many factors as there are eigenvalues greater than 1. Unfortunately, although easy to use, this procedure has often been found to perform poorly.

Scree Plot

Another popular method for determining the number of common factors is the scree plot. The scree plot is a graph of the eigenvalues, plotted from largest to smallest. This graph is then examined to determine where the last major drop in eigenvalues occurs. The number of factors equivalent to the number of eigenvalues that precede the last major drop are retained. For example, in Figure 1, the scree plot would suggest retention of three common factors. Although somewhat subjective, this procedure has been found to function reasonably well when clear dominant factors are present.

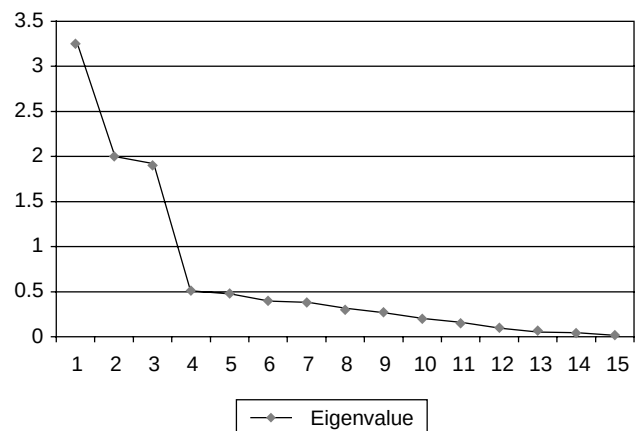


Figure 1 Scree Plot

Parallel Analysis

Parallel analysis involves comparing eigenvalues obtained from the data with eigenvalues that would be expected from random data with an equivalent number of variables and equivalent sample size. The number of factors retained is equivalent to the number of eigenvalues in the sample data set that are greater than the corresponding eigenvalues expected from the random data. Table 4 provides an example of a parallel analysis that suggests retaining three factors. Studies have suggested that parallel analysis functions well when strong factors are present in the data.

Table 4 Sample Parallel Analysis

	<i>Observed Eigenvalues</i>	<i>Eigenvalues Expected From Random Data Set</i>
1.	1.8934	1.3848
2.	1.7839	1.2390
3.	1.7738	1.1907
4.	1.0003	1.0332
5.	.6987	.7986
6.	.4205	.5334
7.	.4133	.4202
8.	.2454	.3454

Goodness of Fit

When conducting certain kinds of EFAs, goodness-of-fit indices can be computed. These are numerical indices that evaluate how well the model accounts for the data. These indices can be compared for a series of models with increasing numbers of common factors. The appropriate number of factors is determined by finding the model in which a model with one less factor demonstrates substantially poorer fit and a model with one more factor provides little improvement in fit.

Model Fitting

The goal of model fitting (also called factor extraction) is to obtain estimates for the model parameters. A variety of methods can be used to accomplish this goal.

Principal Axis Factors

For any EFA model, numerical values can be calculated for the model and used to generate a predicted correlation matrix (i.e., the model's predictions of how the measured variables should be correlated). The predicted matrix can then be compared with the observed correlation matrix to obtain a residual matrix (i.e., a matrix of the differences between the predicted and observed correlations). Noniterated principal axis factors (PAF) compute model parameters such that the sum of the squared residual matrix values is minimized. Iterated PAF uses the same mathematical procedure but with an additional series of steps to refine the estimates. At each step, the estimates from the prior step serve as the starting point for the next set of calculations. This process continues until the estimates at the start of the calculations are extremely similar to the estimates at the end of a step of calculations.

Maximum Likelihood

Maximum likelihood (ML) is a model-fitting procedure based on the likelihood function. The *likelihood function* refers to the relative likelihood that a given model with a set of estimates could have produced the observed data. ML seeks to find the set of estimates for a given model that is maximally likely to have produced the data. One strength of ML is that it provides indices of model fit and confidence intervals for estimates. A disadvantage of ML is that, unlike PAF, it assumes the data are multivariate normal.

Principal Components Analysis

A final model-fitting procedure that is sometimes used for EFA is principal components analysis (PCA). Although PCA is popular and computationally similar in some respects to PAF, this method is not a factor analysis in the strict sense of the term. Specifically, the method is not based on the common factor model. Most notably, this procedure does not distinguish between common and unique variance. Thus, PCA does not take into account the existence of random error.

Rotating a Solution

When examining EFA models with more than one factor, there will be a family of best-fitting solutions for the data. Thus, it is necessary to choose which of these equally fitting solutions is most readily interpretable. This process is accomplished using rotation. Most rotations seek the solution with the best “simple structure.” According to Louis Thurstone, simple structure exists when each factor influences a distinct subset of measured variables, there is little overlap in the subsets of measured variables, and each measured variable is influenced only by a subset of the common factors.

Numerous rotations have been proposed. Some of these procedures are *orthogonal rotations* (i.e., rotations that assume factors are uncorrelated). The most widely used orthogonal rotation is Varimax rotation. Other rotations are *oblique rotations* (i.e., rotations that allow, but do not require, factors to be correlated). Popular oblique rotations include Direct Quartimin, Promax, and Harris-Kaiser Orthoblique rotation. Orthogonal and oblique rotations will lead to similar results if factors are relatively uncorrelated, but oblique rotations may produce better simple structure when factors are correlated.

—Naomi Grant and Leandre Fabrigar

See also Exploratory Data Analysis; Factor Analysis; Factor Scores; Multiple Factor Analysis

Further Reading

- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21, 230–258.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

SAS and SPSS syntax files to assist in exploratory factor analysis: <http://flash.lakeheadu.ca/~boconno2/nfactors.html>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, 6(2), 147–168.

When there are large data sets, factor analysis is often selected as the technique to help reduce the data set to a more manageable size (and to see how well this data reduction fits on the hypothesis that may have been proposed). In this study, the authors surveyed **exploratory factor analysis** (EFA) practices in three different organizational journals, published from 1985 through 1999, to investigate the use of EFA. The review of 371 studies shows reason for optimism, with the tendency to use multiple number-of-factors criteria and oblique rotations increasing. The authors also found that researchers tend to make better decisions when EFA plays a more consequential role in the research. They stress the importance of careful and thoughtful analysis, including decisions about whether and how EFA should be used.

EYEBALL ESTIMATION

Eyeball estimation refers to inspecting data and quickly making an educated guess about the approximate magnitude of relevant statistics without using a calculator or statistical tables.

Here are some examples:

- To eyeball estimate the mean from data presented as a histogram, imagine that the histogram is cut out of plywood. The mean of the distribution is the point where that piece of plywood would balance.
- To eyeball estimate the mean from data presented in a table, find the largest and the smallest values; the mean will be approximately halfway between those values. For example, if the values of *X* are 7, 8, 6, 7, 5, 6, 4, 7, 6, 8, 9, 7, 8, 9, 7, 6, 6, 7, 4, 5, 6, 5, 4, 8, 7, the largest value is 9 and the smallest value is 4; the mean will be approximately 6.5. (It is actually 6.48.)

- To eyeball estimate the standard deviation from data presented as a histogram, superimpose a sketch of a normal distribution over the histogram—make the normal distribution cut through the tops of some of the histogram’s bars. The standard deviation will be approximately the distance from the mean to the inflection point of the normal distribution.

- To eyeball estimate the standard deviation from data presented in a table, find the range (the largest value minus the smallest value). The standard deviation is roughly a quarter of the range. For example, for the data above, the largest value is 9 and the smallest value is 4, so the range is $9 - 4 = 5$. The standard deviation will be approximately $5/4$, or 1.25. (It is actually 1.45.)

The other commonly used descriptive statistics (correlation coefficients, regression constants, areas under normal distributions) can also be eyeball estimated, as can straightforward inferential statistics such as *t* tests and analyses of variance.

Eyeball estimation is not a substitute for accurate computation. Eyeball estimates are “in-the-ballpark” approximations that can be affected (sometimes dramatically) by such factors as skew. However, eyeball estimation is a valuable skill. It enables the observer

to get a sense of the data and to spot mistakes in the computations.

Students benefit from eyeball estimation because it cultivates genuine understanding of statistical concepts. The ability to make an in-the-ballpark, educated guess is better evidence of comprehension than is the ability to compute an exact result from a formula. Furthermore, eyeball estimation is quick. A beginning student can eyeball estimate a standard deviation in about 15 seconds; computation would take the same student about 15 minutes. Furthermore, while students are eyeball estimating standard deviations, they are developing their comprehension of the standard deviation as a measure of the width of a distribution. By contrast, almost no time involved in the computation of a standard deviation is focused on that comprehension.

—Russell T. Hurlburt

See also Bar Chart; Line Chart; Pie Chart

Further Reading

Hurlburt, R. T. (2006). *Comprehending behavioral statistics*. Belmont, CA: Wadsworth.

F

There is a very easy way to return from a casino with a small fortune: go there with a large one.

—Jack Yelton

FACE VALIDITY

Face validity is most often understood as a subjective and cursory judgment of a concept, assessment instrument, or any other conceptualization to ascertain whether, on its face, it appears valid (i.e., that the concept being measured seems reasonable; that a test instrument appears to measure what it purports to measure; that the association between the concept and how it is measured seems appropriate and relevant at first glance), without further regard to the underlying legitimacy of the nomological network, concept, instrument and test items, or the construct it purports to measure.

Face validity is the least reliable validity judgment among validity measures and should serve only as a preliminary screening, given that it addresses appropriateness without empirical data. However, if the minimum requirement of face validity cannot be established, then it is highly unlikely that any of the more rigorous validity criteria will hold.

Face validity, therefore, may reflect reasonable, consistent, and understandable surface connections between the instrument and test items on the one hand and their underlying construct on the other. Conversely, face validity might fail to reveal such

connections, irrespective of whether, on closer scrutiny, they actually exist. Therefore, there can be no claim of a logical relationship between face validity and true validity, although correlations between face validity and true validity are possible.

Judgments about face validity are closely connected to the knowledge and experience of the test user. For example, the more a test instrument and its items appear to test takers, on the basis of their experience, to be understandable, reasonable, and clearly related to the test criterion, the more likely it will be that the test takers will judge the test to have a high level of face validity. However, the face validity of a test instrument is more likely to be judged accurately by a psychometrician than by an individual without psychometric training.

Practical Aspects

Advantages

The notion of face validity embodies a number of advantages; for example, it may enable someone to narrow the number of instruments or reports under consideration. However, its most significant contribution is bringing the experience and contexts of test takers into consideration during test construction. For example, potential test items might be examined

by representatives of potential test populations to determine whether the items are recognizable and appropriate for inclusion in the final version of a test. That is, the level of face validity is established by a test taker who rates a test, test item, or battery of tests as relevant or irrelevant. Face validity, therefore, is judged from the point of view of the user's knowledge, experience, and attention.

If test items appear to be related to appropriate and relevant content, test takers are more likely to be highly motivated to do their best, thereby making the instrument more valid and reliable. Equally, if the content of the test or items is perceived to be inappropriate or irrelevant, then the test takers' level of motivation for test performance may well be lower. Higher levels of test-taking motivation ensure that test takers are also likely to be more cooperative, more satisfied with their performance, and less likely to blame the test if they obtain low scores.

Relatedly, high levels of face validity are significant when selecting tests that rely heavily on test takers' cooperation. Finally, high levels of face validity may also be a potent factor in marketing the test commercially.

Disadvantages

Face validity also comprises a number of less positive attributes. For example, face validity may actually, in some instances, work against high levels of motivation to complete a test properly if test takers are coerced in some way to take the test and if they are not accustomed to test-taking behavior. In some test areas, such as tests that measure levels of performance rather than more abstract concepts, any distortion related to this motivational attribute may not be significant. However, among tests that attempt to measure more subjective areas (such as personality traits), a test with high face validity might well fail because respondents do not wish to answer the test items truthfully.

Other problems might occur. For example, while high levels of motivation on the part of test takers may be assumed and desirable, there is the possibility that even their most diligent attempts might not reflect

what the test is actually measuring. Further, the only evidence of the test or test items' face validity is the judgment of the test taker.

In sum, face validity is considered the chief means of generating acceptance from the general public, from organizations that are considering using a test, and from test takers themselves.

Context

Historically, the exact nature of validity in general has undergone several changes and has assumed great significance in the area of psychometrics.

While commonly acknowledged, by implication, as tangential in test and test item construction for many decades, issues of face validity were largely ceded to the publishers and other vendors of psychometric tests by the early 1900s. Face validity emerged more prominently in the 1940s and 1950s, but dissension among scholars caused several leaders in the field to suggest that use of the term *face validity* be eliminated from psychometrics altogether.

This problem was largely put to rest by the publication of the first set of *Standards for Educational and Psychological Tests* by the American Psychological Association in 1954. The *Standards* largely reinforced the claims made for face validity by Lee Cronbach (1949) and colleagues, particularly Anne Anastasi (1954). Proponents of face validity claimed (a) that it was distinct in its own right, (b) that it was different from other forms of validity, (c) that face validity was not interchangeable with other forms of validity, and (d) perhaps most important, that it was a key practical aspect of test construction related to test acceptability in the eyes of consumers and test takers.

Subsequent American Psychological Association standards have, however, increasingly downplayed the worth of face validity. The 1966 and 1974 *Standards* specified only that face validity should, in psychometrics, be separated from more substantial forms of content validity in psychometrics. The 1985 *Standards* ignored the issue of face validity altogether. That same year, however, Baruch Nevo made the case that face validity should be considered and reported in test construction, although primarily to enhance

acceptability rather than as a formal psychometric aspect of test construction.

Implied Meanings of Face Validity

Nevo argued that three and possibly four meanings of face validity have a bearing on judgments about instruments and their test items: validity by assumption, by definition, by appearance, and possibly by hypothesis.

Assumption

Face validity is established when the test user assumes the predictability of the tested criterion by identifying the degree of reasonableness of the test items as related to the objective of the test. The assumption thus made is so strong that recourse to further statistical evidence is unlikely or unwarranted. Establishing face validity in this way is problematic because whatever the level of the test's or test items' perceived practical value, more substantial statistical evidence, whether supportive or conflicting, may be disregarded.

Definition

Face validity by definition makes a judgment of tests or test items via a sample judged by an expert to thoroughly represent the universe of such questions. Historically, this meaning comes closest to what face validity was intended to explain. The better defined the test criterion and the closer it is related to the test items themselves, the more likely that face validity, by definition, can be established. Obviously, face validity assumptions can only be extrapolated to the larger population from which the sample items were drawn.

Appearance

Face validity by appearance makes judgments of a test instrument and its items without recourse to statistical tests to verify stability. Validity is established by those who judge the test and its items relevant, practical, and closely related to the purpose of the test and test performance criteria. Such tests, therefore, are

likely to have a high degree of acceptance among those who use them as well as those who take the tests.

Hypothesis

Face validity by hypothesis is arguably a secondary consideration associated with assumption and definition. Face validity is judged in this case when it is necessary and practical to use a test in the real world before statistical information can validate the test. The test is hypothesized as having at least some degree of validity on the basis of other valid tests with the same or similar test criteria and objectives. This form of face validity differs from the first three, all of which judge face validity on easily identifiable and logical ties between the test items and the test criterion. With face validity by hypothesis, the level of confidence in the test's validity rests on the level of confidence in the hypothesis and the amount of research that supports it. Furthermore, the level of confidence in the hypothesis will determine the feasibility of when, how, or whether the testing should proceed.

Other Aspects

Since 1985, discussion of face validity has not been widespread, although some isolated pockets of interest seem to persist. For example, Mark Mostert constructed and applied face validity criteria to meta-analyses in special education, contending that meta-analysis is often assumed to derive definitive quantitative answers from an entire body of research. However, face validity of published meta-analyses (in special education, in this case) can be substantially affected by the information supplied to the user, an observation that has important implications for theory and practice.

This study noted that published meta-analytic results rely heavily on several essential interpretive aspects, including (a) the definition and relationships between the primary study independent variables, (b) the manner in which the independent variables are coded, and (c) how these key variables are interrelated and reported. Face validity is especially germane in view of meta-analyses that have been conducted

on the same body of primary studies but that have yielded dissimilar findings.

To establish the exact nature of face validity in special education meta-analyses, the study developed a set of criteria to clarify mega-analytical study characteristics that needed to be available in order for a user to judge face validity. The criteria, which encompass six domains, are discussed in the following sections.

Locating Studies and Establishing Context

The first set of information to provide to users in order to establish face validity includes (a) a literature review (to briefly describe studies and to contextualize the meta-analysis), (b) search procedures used to obtain the primary studies, (c) the dates of the search, (d) the number of primary studies used in the meta-analysis (to establish whether they are a population or a sample of a known universe of studies), and (e) confirmation that the primary studies are clearly noted.

Specifying Inclusion Criteria

The primary study data set must also be justified by reporting the criteria used to select the primary studies and the criteria used to eliminate other primary studies.

Coding Independent Variables

In this step, the meta-analyst must provide (a) a general description of the primary studies around the central research question, (b) a description of the independent variables, (c) descriptions of relationships between variables to explain the conceptual and rational connections between variables (if more than one variable is to be entered into the meta-analysis), and (d) notes explaining any variation among the coded variables.

Calculating Individual Study Outcomes

The meta-analysis requires extensive reporting of the statistical calculations, including (a) the number of effects sizes (ESs) calculated; (b) ES range and standard deviation as general indicators of the scope of

variability found in the primary studies, noting both n sizes (ESs for each primary study) and the overall N (used to calculate the ES for the meta-analysis) to measure the overall effect of the meta-analyzed intervention; (c) factors affecting ES (e.g., pooled ESs or the use of placebo groups); and (d) interrater reliability to demonstrate coding of the independent variables by more than one researcher in order to add credence to the analysis and the overall interpretation of the meta-analytical outcomes.

Analyzing Data

After executing and reporting the basic statistical calculations, the analyst should proceed to add interpretive aspects: (a) reporting fail-safe sample size (the number of nonsignificant studies needed outside of those in the meta-analysis to negate the meta-analytic results); (b) summarizing statistics for significant findings (e.g., F and t ratios or r s; useful for drawing generalized research conclusions); (c) reporting nonsignificant findings along with or instead of significant findings to establish the overall integrity of the analysis; (d) explaining the proportion of variance accounted for by the treatment effect after statistical artifacts and other moderators have been acknowledged; (e) providing a summation of research applications and important findings of the meta-analysis, adding analytical coherence to the research hypothesis; and (f) suggesting how findings may be practically and theoretically applied.

Documenting the Limits of the Meta-Analysis

Finally, the limits of the meta-analytic findings should be discussed in order to circumscribe the interpretation of the data.

On the basis of these face validity criteria, the study reported that of 44 special education meta-analyses, the mean proportion of face validity criteria evident from the publications was .60, with a range of .26–1.0.

—Mark Mostert

See also Psychometrics; Validity Theory

Further Reading

- Adams, S. (1950). Does face validity exist? *Educational and Psychological Measurement*, 10, 320–328.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Gynther, M. D., Burkhart, B., & Hovanitz, C. (1979). Do face validity items have more predictive validity than subtle items? *Journal of Consulting and Clinical Psychology*, 47, 295–300.
- Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, 10, 93–98.
- Mostert, M. P. (2001). Characteristics of meta-analyses reported in mental retardation, learning disabilities, and emotional and behavioral disorders. *Exceptionality*, 9(4), 199–225.
- Turner, S. P. (1979). The concept of face validity. *Quality and Quantity*, 13, 85–90.

FACTOR ANALYSIS

Factor analysis (FA) is a statistical technique used to examine the structure of correlations among a set of observed scores. Although initially developed as a method for studying intelligence, FA has become widely used in numerous areas of psychology as well as in other social sciences, business, economics, and biology.

The Common Factor Model

Although Charles Spearman is largely credited with development of the first formal FA model, Louis L. Thurstone is generally regarded as having provided the foundations of contemporary FA with his development of the common factor model (CFM). The goal of the CFM is to represent the structure of correlations among observed scores by estimating the pattern of relationships between the common factors and the measured variables (a set of observed scores). This relationship is represented numerically by a factor loading in the analysis. Thurstone posited that each measured variable in a set of measured variables is a linear function of two types of latent factors. Common factors, which are unobserved latent variables (constructs), influence

more than one measured variable and thus account for the correlations among the measured variables. Unique factors, also latent variables, influence only one measured variable in a set and thus do not account for correlations among measured variables. Unique factors consist of two components (i.e., a specific factor and an error of measurement) although in practice these components are not separately estimated.

The CFM can be expressed in several forms. When expressed as a data model, its goal is to explain the structure of the raw data. Each participant's score on each of the measured variables is represented by a separate equation. The data model for the CFM is expressed by the equation

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \dots + \lambda_{jm}z_{im} + v_{ij},$$

where

i is the individual,

j is the measured variable,

x_{ij} is the score of the individual on the measured variable,

μ_j is the mean of the measured variable,

z_{im} is the common factor (latent variable) score for individual i on factor m ,

λ_{jm} is the factor loading on test j on factor m , and

v_{ij} is the unique factor score for person i on the unique factor j .

The components of the unique factor are represented by the equation

$$v_{ij} = s_{ij} + e_{ij},$$

where

s_{ij} is the factor score of individual i on the specific factor j and

e_{ij} is the factor score of individual i on the unique factor j .

Although the data model is conceptually useful, in practice its values are impossible to estimate because individuals' scores on unobservable latent variables cannot be known.

The data model, however, does provide the theoretical basis for an alternative expression of the model that can be used. Specifically, given the data model and making certain distributional assumptions, it is possible to mathematically derive a version of the CFM designed not to explain the structure of raw data but instead to account for the structure of correlations among measured variables. The correlational structure version of the CFM is represented by the equation

$$P = \lambda \phi \lambda^T + D_\psi,$$

where

P is the correlation matrix in the population,

λ is the factor loadings matrix,

ϕ is the matrix of correlations among common factors,

λ^T is transpose of the factor loadings matrix, and

D_ψ is the matrix of unique variances.

This equation states that a matrix of correlations among measured variables is a function of the common factor loadings (λ), the correlations among common factors (ϕ), and the unique variances (D_ψ).

Yet another way in which the CFM can be represented is in the form of a “path diagram” (see Figure 1). In these diagrams, circles or ovals represent latent variables and factors (both common and unique), and squares or rectangles represent measured variables. The other components of the diagram are directional arrows, which imply a linear causal influence, and bidirectional arrows, which represent an association with no assumption of causality. Figure 1 provides an example of how the CFM can be represented in a case where it is hypothesized that two common factors (F1 and F2) and four unique factors (U1–U4) can be used to explain correlations among four measured variables (X1–X4). Note that in the present example, the two common factors are assumed to each influence two measured variables, and each unique variable influences only one measured variable.

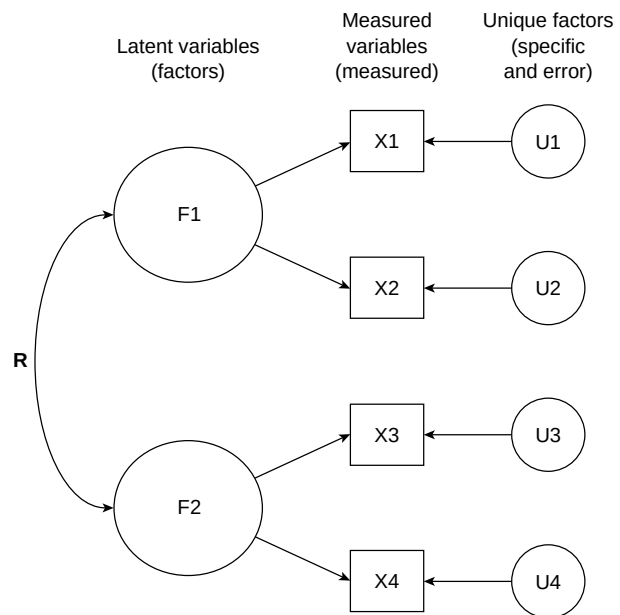


Figure 1 Path Diagram for a Common Factor Model With Two Common Factors and Four Measured Variables

Exploratory Factor Analysis

In practice, factor analysis involves a set of statistical procedures used to arrive at numerical values for the elements of the matrices expressed in the correlation structure version of the CFM. In some cases, this is accomplished in an exploratory fashion. Exploratory factor analysis (EFA) (also called unrestricted factor analysis) is used when there is little empirical or theoretical basis for specifying a precise a priori model. In brief, EFA is a collection of procedures for determining the precise model that is appropriate for the data and arriving at estimates of the numerical values for that model. Several steps must be undertaken to accomplish these objectives.

First, the correct number of common factors must be determined. Procedures for performing this step include the scree plot, parallel analysis, model fit, and Kaiser criterion (the “eigenvalue is greater than one” rule). Importantly, not all these procedures work well (e.g., the Kaiser criterion), and none are infallible. Thus, methodologists recommend using several of the best-performing procedures in conjunction with one another. Methodologists also recommend that

researchers consider the interpretability and replicability of the solution in addition to these statistical procedures when determining the appropriate number of factors.

Second, given a model with a specific number of factors, the model must then be fit to the data (a process also called factor extraction). This process involves calculating the actual numerical values for the model. A number of model fitting procedures are available (e.g., noniterated principal axis factors, iterated principal axis factors, and maximum likelihood). Although the procedures use different mathematical approaches, all of them share a common goal of attempting to find the set of numerical values that will result in the best fit of the model to the data. One advantage of the principal axis factors methods is that they do not make strong distributional assumptions. In contrast, maximum likelihood does make assumptions of multivariate normality but provides more information (e.g., model fit indices, confidence intervals for parameter estimates).

Finally, because more than one best-fitting solution is possible for an EFA involving two or more factors, a single solution is arrived at through “rotation.” The goal of the rotation process is to select that solution that is most readily interpretable. Various rotation procedures have been proposed. Some of these procedures (e.g., varimax rotation) are orthogonal rotations in that they assume common factors to be uncorrelated. Others (e.g., direct quartimin, Promax, and Harris-Kaiser orthoblique) are oblique rotations in that they permit (but do not require) common factors to be correlated. Given that in most cases it is difficult to know if factors will be correlated, oblique rotation is usually more conceptually sensible than orthogonal rotation.

Confirmatory Factor Analysis

In other cases, researchers may have a strong empirical or theoretical basis to make predictions regarding the number and nature of the common factors. In these contexts, confirmatory factor analysis (CFA) is used. CFA can be broken into several phases. First, the

researcher must specify the model. This process involves specifying how many common factors exist and exactly where zero factor loadings will occur (i.e., which measured variables will not load on each common factor). Model specification also requires the researcher to specify which, if any, common factors are correlated with one another and which, if any, unique factors will correlate with one another.

Once the model is specified, it must be fit to the data. As in EFA, this involves finding the set of numerical values that provides the best fit of the model to the data. Various model fitting procedures are available (e.g., generalized least squares and asymptotic distribution-free estimation), but maximum likelihood is by far the most frequently used procedure. Because the specific pattern of zero and nonzero factor loadings is specified in CFA, rotation of solutions is not necessary.

The third phase of CFA is model evaluation. This step involves examining the results of the analysis and assessing the adequacy of the proposed model. Several types of information are considered. For example, most CFA model fitting procedures permit the computation of goodness-of-fit indices. Numerous fit indices have been proposed, but these indices are often categorized as falling into two groups. Absolute fit indices assess the absolute magnitude of the discrepancy between the model and the data. Popular indices of this type include root mean square error of approximation and standardized root mean square residual. Incremental fit indices evaluate the fit of the model to the data relative to some comparison model (usually the null model, which postulates no underlying structure in the data). Popular fit indices of this type include the nonnormed fit index (or Tucker-Lewis fit index) and the normed fit index.

The second category of information used in model evaluation is the parameter estimates of the model. Unlike EFA, CFA analyses not only report the estimates of the parameters but also routinely report confidence intervals and significance tests for all estimates. All of this information is examined to evaluate the theoretical plausibility of the estimates. Additionally, it is sometimes the case in CFA that a

researcher has specific hypotheses to test regarding certain parameters in the model. For example, the researcher might postulate that a given factor is more highly correlated with one factor than another. Precise hypotheses involving comparisons of parameters can be tested by placing equality constraints on the model (e.g., constraining the two correlations being compared to be equal) and then comparing the constrained model to the original model. A formal statistical test comparing the fit between the two models (a chi-square difference test) can then be conducted. If significant, such a test indicates that the constraint is not appropriate and thus the constrained parameters are significantly different from one another.

The final step in CFA is model modification. When a proposed model is found to perform poorly (as a result of either poor model fit or implausible parameter estimates), researchers sometimes consider modifications to the model. Most structural equation modeling programs used to conduct CFA provide numerical indices that can be used as a guide to which parameters originally fixed in the model might be freed to improve model fit. Unfortunately, use of these indices has proven problematic because such changes often do not have a firm theoretical basis. Additionally, studies have suggested that these modification indices are not especially effective in identifying model misspecifications. Thus, most methodologists recommend that model modification be guided by theory rather than the use of these indices.

Conclusions

EFA and CFA can be thought of as complementary rather than opposing approaches to data analysis. Specifically, EFA may be the approach of choice during the early phases of a research program, when comparatively little is known about the underlying structure of correlations among a set of measured variables. Once EFA analyses have helped establish a firm theoretical and empirical basis for more precise predictions, later studies may make use of CFA to conduct more rigorous and focused tests of the researchers' theory.

—Ronald D. Porter and Leandre R. Fabrigar

See also Exploratory Factor Analysis; Factor Scores; Multiple Factor Analysis

Further Reading

- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, 31, 439–485.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale NJ: Erlbaum.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Society.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Wegener, D. T., & Fabrigar, L. R. (2000). Analysis and design for nonexperimental data: Addressing causal and noncausal hypotheses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412–450). Cambridge, UK: Cambridge University Press.

Factor analysis: http://en.wikipedia.org/wiki/Factor_analysis

Factor analysis software sampling: <http://quantrm2.psy.ohio-state.edu/browne/software.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65(2), 202–226.

Factor analysis is one of many techniques that allows us to better understand threads common to large sets for data. The purpose of this study was to investigate the relationship between sample size and the quality of factor solutions obtained from one type of factor analysis, exploratory factor analysis. This research expanded on the range of conditions previously examined, employing a broad selection of criteria for the evaluation of the quality of sample factor solutions. Results

showed that when communalities are high, sample size tended to have less influence on the quality of factor solutions than when communalities are low. Overdetermination of factors was also shown to improve the factor analysis solution. Finally, decisions about the quality of the factor solution depended on which criteria were examined.

FACTOR SCORES

The purpose of conducting factor analysis is to explain the interrelationships between a set of measured variables and a reduced set of theoretically meaningful common factors. An attractive feature of this multivariate procedure is the capability to rank order the measured objects (usually people) on the common factors. These novel factor scores can be used in subsequent statistical analyses or employed in decision making processes. For instance, factor scores can be correlated with other variables, entered as predictor variables in regression analyses, or used as dependent measures in analyses of variance. They can also be employed in applied settings, such as when a clinical psychologist uses a client's factor scores on measures of psychological well-being to determine a treatment plan or when a school psychologist uses factor scores from an intelligence test to make judgments regarding a child's cognitive abilities. Given their utility, factor scores are widely employed in both research and practice.

Factor Scores Explained

As a contrived example study, consider 200 individuals who rate themselves on six questionnaire items written to measure personality traits. The individuals rate on a 5-point scale the extent to which each statement (e.g., "I have many close friends," "I do not get stressed-out easily") applies to themselves. A common factor analysis is subsequently conducted on the ratings, and two factors are extracted. After the factors are rotated with an oblique transformation, they are labeled Extroversion and Emotional Stability. Factor

scores for the 200 individuals can now be computed by regressing the six item scores onto the two factors. Common factors are often referred to as *latent* or *unobservable* because their scores must be derived through such a regression analysis based on the original items. The resulting regression weights are referred to as *factor score coefficients*, and they can be applied to the standardized item responses to compute the factor scores. For example, Extroversion factor scores may be computed as follows:

$$\begin{aligned} &(\text{Item1}_z)(.45) + (\text{Item2}_z)(.32) + (\text{Item3}_z)(.72) \\ &+ (\text{Item4}_z)(.02) + (\text{Item5}_z)(-.12) + (\text{Item6}_z)(.05). \end{aligned}$$

The values .45, .32, .72, and the rest are the factor score coefficients, which are standardized regression weights. Coefficients are computed for all six items, and their relative absolute magnitudes indicate that the first three items contribute most to the prediction of scores on the Extroversion factor, while the remaining three items contribute less (their weights are near zero). Item 5 contributes negatively to the computation of Extroversion factor scores in this example.

Now consider two individuals, Joe and Mary. If their standardized responses (i.e., their *z* scores) on the rating scale are placed in the equation, their Extroversion factor scores are as follows:

$$\begin{aligned} \text{Joe's Extroversion} &= (.90)(.45) + (1.02)(.32) + (1.10) \\ &(.72) + (.25)(.02) + (.43)(-.12) \\ &+ (.22)(.05) = 1.49 \end{aligned}$$

$$\begin{aligned} \text{Mary's Extroversion} &= (-.68)(.45) + (.19)(.32) \\ &+ (-1.29)(.72) + (.45)(.02) \\ &+ (.77)(-.12) + (.15)(.05) \\ &= -1.25 \end{aligned}$$

Since high scores on the rating scale indicate greater extroversion, Joe is found to be extroverted and Mary is introverted. The standardized Emotional Stability factor scores can be computed similarly:

$$\begin{aligned} \text{Joe's Emotional} &= (.90)(.05) + (1.02)(.01) + (1.10) \\ \text{Stability} &(-.03) + (.25)(.60) + (.43)(.70) \\ &+ (.22)(.53) = .59 \end{aligned}$$

$$\begin{aligned} \text{Mary's Emotional} &= (-.68)(.05) + (.19)(.01) \\ \text{Stability} &+ (-1.29)(-.03) + (.45)(.60) \\ &+ (.77)(.70) + (.15)(.53) = .90 \end{aligned}$$

The second numbers (e.g., .05, .01, -.03) in each product are the factor score coefficients for the Emotional Stability factor, and the first numbers are Joe's and Mary's *z* scores for the six items. The results indicate that Mary is slightly more emotionally stable than Joe. The same coefficients for both factors could similarly be used to compute factor scores for the remaining 198 individuals in the study.

Computing factor scores is thus a matter of regressing the items onto the factors to derive factor score coefficients. These standardized regression weights are then applied to the standardized scale responses to compute each individual's relative standing on the factors. How, exactly, are the factor score coefficients computed? Unfortunately, the answer to this question is not straightforward because of a historic intellectual puzzle referred to as *factor score indeterminacy*.

In the early days of factor analysis, a peculiar property of the common factor model was discovered; specifically, the algebra underlying the computation of factor scores involves solving a set of equations with more unknowns than equations. The implication is that a unique solution for the factor scores does not exist. Instead, an infinite number of sets of factor scores can be computed for any given analysis. The scores for Joe and Mary above on the Extroversion and Emotional Stability factors are thus one of an infinite number of sets of factor scores that can be computed! What would happen if in another solution, Joe was introverted and Mary was extraverted? There would be no way of choosing between the two different conclusions regarding Joe and Mary since both sets of factor scores represent valid solutions. Obviously, factor score indeterminacy is a serious issue. If the individuals (or measured

objects) in the study cannot be unambiguously rank ordered along the Extroversion and Emotional Stability factors, one must wonder if the factors are of any scientific value.

Refined and Coarse Factor Scores

Since it was discovered in the early 1900s, psychometricians have attempted to understand the analytical, practical, and theoretical meaning of factor score indeterminacy. One implication is that computed factor scores are imperfect representations of the factors themselves. Consequently, the statistical properties of the factor scores may not always match those of the factors for which they are computed. For instance, while the common factors in a given analysis are uncorrelated (i.e., orthogonal), the factor scores may be slightly or moderately correlated. To adjust for these discrepancies, psychometricians have developed a number of methods for computing factor scores, and these methods can be divided into two groups: *refined* and *coarse*. The former produce multidecimal factor score coefficients like those in the examples above. They are often referred to by their distinct statistical properties or by the names of the scholars responsible for their derivation and are summarized in Table 1.

Table 1 Methods for Computing Refined Factor Scores

<i>Author</i>	<i>Name</i>	<i>Desirable Property</i>
Thurstone	Regression	Indeterminacy is minimized (i.e., determinacy is maximized).
Anderson-Rubin	Orthogonal	Factor scores for different factors are uncorrelated (i.e., perfectly orthogonal); only appropriate for orthogonal factors.
Bartlett	Univocal	Factor scores for a given factor are not correlated with other factors in the analysis; only appropriate for orthogonal factors.
McDonald	Correlation-preserving	Correlations between factor scores match the correlations between the factors; appropriate for orthogonal or correlated factors.

The most popular of the refined factor scoring methods can be found in many statistical programs, and their equations are readily available (see Further Reading at the end of this entry). None of these methods are clearly superior to the others, however, and the choice between the different refined factor scores rests entirely with the researcher, who alone can decide which property is most important for a particular investigation.

The coarse methods offer popular alternatives to the refined factor scores. With coarse methods the factor scores are computed as simple sums of the original or standardized item scores. For example, Joe's Extroversion factor score might be computed as follows:

$$\begin{aligned}\text{Joe's Extroversion} &= \text{Item1}_z + \text{Item2}_z + \text{Item3}_z \\ &= .90 + 1.02 + 1.10 = 3.02\end{aligned}$$

Such scores are often referred to as scale scores, sum scores, or total scores, among other names. As long as they are based on the results of a factor analysis, they are also appropriately referred to as factor scores since they indicate the relative standings of the people (or objects) on the factors. This point can be understood by realizing that the simple summing process involves an implicit weighting scheme:

$$\begin{aligned}\text{Joe's Extroversion} &= (1)(\text{Item1}_z) + (1)(\text{Item2}_z) \\ &\quad + (1)(\text{Item3}_z) + (0)(\text{Item4}_z) \\ &\quad + (0)(\text{Item5}_z) + (0)(\text{Item6}_z) \\ &= (1)(.90) + (1)(1.02) + (1)(1.10) + (0)(.25) \\ &\quad + (0)(.43) + (0)(.22) = 3.02\end{aligned}$$

It is easy to see that the coarse method of computing Joe's factor scores is a simplification of the refined method; the multidecimal factor score coefficients have been replaced with whole numbers. In this example, the fourth, fifth, and sixth items are given weights of zero, essentially dropping them from the computation of the coarse factor scores. Since the six items are all measured on the same scale, it should also be noted that Joe's coarse factor scores could be computed by summing his original responses instead of the z scores.

The coarse methods differ with regard to the processes used to determine the whole weights. The most common approach for determining the weights is to examine the structure coefficients (i.e., the correlations between the factors and the items) and choose the most salient items. A salience criterion is determined using some rule-of-thumb value such as $|.30|$ or $|.40|$. Items with salient positive structure coefficients are given weights of 1, and items with salient negative coefficients are given weights of -1 . Items with non-salient structure coefficients are given weights of zero. The pattern coefficients are sometimes used instead of the structure coefficients in this scheme. Either way, these particular coarse factor scores are problematic because they run the risk of being very poor representations of the factors themselves. In other words, coarse factor scores based on the structure or pattern coefficients may not be highly correlated with the very factors they are intended to quantify. Consequently, although they are popular in the scientific literature, they are not recommended. Instead, the coarse factor scores should be based on a process of simplifying the factor score coefficients. The process therefore begins by using one of the refined factor scoring methods to derive the factor score coefficients. The resulting coefficients are then simplified to whole numbers and used as weights to compute the coarse factor scores. This coarse method is superior because it is based on the factor score coefficients, which are specifically designed for computing factor scores. The relative magnitudes of the pattern and structure coefficients may be quite discrepant from the factor score coefficients, leading to an inaccurate weighting scheme and invalid factor scores.

Evaluating Factor Scores

Tools are available for assessing the amount of indeterminacy in any given common factor analysis as well as for evaluating the quality of the computed factor scores. Many statistical programs, for instance, report the squared multiple correlations for the common factors. These values indicate the proportion of determinacy in the common factors, and results near 1 are desirable ($1 = \text{no indeterminacy}$), whereas results equal

to or less than .80 are generally considered to indicate too much indeterminacy. When a common factor is judged as highly indeterminate, factor scores should not be computed via any method above. If a factor is judged sufficiently determinate, and factor scores are computed, they should be evaluated for their validity, correlational accuracy, and univocality. Validity refers to the correlation between the factor scores and the factors themselves, and correlational accuracy refers to the correspondence between the factor score intercorrelations and the factor intercorrelations. Univocality refers to the degree of overlap between factor scores and noncorresponding factors in the analysis. Macros for SAS are readily available for assessing these different properties of refined and coarse factor scores (see Further Reading at the end of this entry).

Additional Issues

A number of additional issues regarding factor scores should be noted. First, principal component analysis (PCA) always produces refined component scores that are determinate. The four methods in Table 1 thus produce identical results for PCA. If coarse component scores are computed, however, they should be based on the component score coefficients and evaluated for their quality. Second, image analysis is the only common factor model that yields determinate factor scores. It is not as popular as other factor techniques (e.g., maximum likelihood or iterated principal axes), but it offers a viable alternative if one wishes to avoid indeterminacy and still use a common factor model. Last, other latent trait statistical methods such as structural equation modeling and item response theory are indeterminate. Although it has not been discussed at great length with these other methods, indeterminacy is an important concern that should not be overlooked indefinitely.

—James W. Grice

See also Exploratory Factor Analysis; Multiple Factor Analysis

Further Reading

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.

Guttman, L. (1955). The determinacy of factor score matrices with applications for five other problems of common factor theory. *British Journal of Statistical Psychology*, 8, 65–82.

Steiger, J. H., & Schonemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). Chicago: University of Chicago Press.

Factor scores computation and evaluation tools: <http://psychology.okstate.edu/faculty/jgrice/factorscores/>

FACTORIAL DESIGN

Factorial design is one of the most popular research tools in many areas, including psychology and education. Factorial design investigates two or more independent variables simultaneously, along with interactions between independent variables. Each independent variable can be either a treatment or a classification. Ideally, all possible combinations of each treatment (or classification) occur together in the design. The purpose is to investigate the effect of each independent variable and interaction on a dependent variable.

In a one-way design, a single independent variable is investigated for its effect on a dependent variable. For example, we might ask whether three therapies produce different recovery rates or whether two drugs lead to a significant difference in average adjustment scores. In a factorial design, we might ask whether the two drugs differ in effectiveness and whether the effectiveness of the drugs changes when they are applied at different dosage levels. The first independent variable is the type of drug (with two levels), and the second independent variable is the dosage (with two levels). This design would be a 2×2 factorial design. Each independent variable could have more than two levels.

An Example of a 2×2 Design

Table 1 presents hypothetical data and means in which the data are presented for scores under two drugs and two dosage levels. Each of the four cell means is based on three observations. That is, three different

Table 1 Summary Table for Two Factor Design

Drug/ Dosage	B_1 10 mg	B_2 20 mg	Overall
A_1	8.0	10.0	
Drug 1	7.0	9.0	
	<u>6.0</u>	<u>8.0</u>	
Means	7.0	9.0	$M_{A_1} = 8.0$
SS	2.0	2.0	$N_{A_1} = 6$
A_2	5.0	3.0	
Drug 2	4.0	2.0	
	<u>3.0</u>	<u>1.0</u>	
Means	4.0	2.0	$M_{A_2} = 3.0$
SS	2.0	2.0	$N_{A_2} = 6$
Overall	$M_{B_1} = 5.5$ $N_{B_1} = 6$	$M_{B_2} = 5.5$ $N_{B_2} = 6$	$M_T = 5.5$

Note: SS = sum of squares; M = mean.

individuals were given either Drug 1 or Drug 2 and either 10 mg or 20 mg of the drug. The 12 individuals were randomly assigned to one of the four combinations of drug and dosage level. Higher mean adjustment scores indicate better adjustment.

As in a simple, independent-sample t test or one-way analysis of variance (ANOVA), the variability within each treatment condition must be determined. The variability within the four groups is defined as SS_{WG} where

$$\begin{aligned} SS_{WG} &= SS_1 + SS_2 + SS_3 + SS_4 \\ SS_{WG} &= 2.0 + 2.0 + 2.0 + 2.0 \\ SS_{WG} &= 8.0. \end{aligned}$$

With three observations in each group, there are $3 - 1 = 2$ degrees of freedom (df) within each group. The degrees of freedom within all groups are

$$df_{WG} = 2 + 2 + 2 + 2 = 8.$$

The within-groups variability is known as the mean square (MS) within groups, defined as

$$MS_{WG} = \frac{SS_{WG}}{df_{WG}} = \frac{8.0}{8} = 1.0.$$

The overall difference between drugs is Factor A, Drug Type. The null and alternative hypotheses are

$$H_0: \mu_{A_1} = \mu_{A_2}$$

$$H_1: \mu_{A_1} \neq \mu_{A_2}.$$

The variability among the drug means is

$$SS_A = N_{A_1}(\bar{X}_{A_1} - \bar{X}_T)^2 + N_{A_2}(\bar{X}_{A_2} - \bar{X}_T)^2$$

$$SS_A = 6(8.0 - 5.5)^2 + 6(3.0 - 5.5)^2$$

$$SS_A = 6(2.5)^2 + 6(-2.5)^2$$

$$SS_A = 6(6.25) + 6(6.25)$$

$$SS_A = 75.$$

With a levels of Factor A, the degrees of freedom will be

$$df_A = a - 1 = 2 - 1 = 1.$$

The MS for Factor A is

$$MS_A = \frac{SS_A}{df_A} = \frac{75}{1} = 75.$$

The F test for Factor A is

$$F_A = \frac{MS_A}{MS_{WG}} = \frac{75}{1} = 75.$$

To test the null hypothesis at $\alpha = .01$, we need the critical value $CV = F_{.99}(1,8) = 11.26$. The null hypothesis is rejected because $F_A = 75.0 > 11.26$. The overall mean, $M_{A_1} = 8.0$, for Drug 1 is significantly greater than the overall mean, $M_{A_2} = 3.0$, for Drug 2. Drug 1 produces significantly greater adjustment than does Drug 2.

Applying similar calculations to Factor B, Dosage Level, we have the hypotheses

$$H_0: \mu_{B_1} = \mu_{B_2}$$

$$H_1: \mu_{B_1} \neq \mu_{B_2}.$$

The variability among the drug means is

$$SS_B = N_{B_1}(\bar{X}_{B_1} - \bar{X}_T)^2 + N_{B_2}(\bar{X}_{B_2} - \bar{X}_T)^2$$

$$SS_B = 6(5.5 - 5.5)^2 + 6(5.5 - 5.5)^2$$

$$SS_B = 0.0.$$

With b levels of Factor B, the degrees of freedom will be

$$df_B = b - 1 = 2 - 1 = 1.$$

The MS for Factor B is

$$MS_B = \frac{SS_B}{df_B} = \frac{0.0}{1} = 0.0.$$

The F test for Factor B is

$$F_B = \frac{MS_A}{MS_{WG}} = \frac{0.0}{1} = 0.0.$$

To test the null hypothesis at $\alpha = .01$, we need the critical value $(CV) = F_{.99}(1,8) = 11.26$. The null hypothesis is not rejected because $F_B = 0.0 < 11.26$. The overall mean, $M_{B_1} = 5.5$, for 10 mg is not significantly different from the overall mean, $M_{B_2} = 5.5$, for 20 mg. There is no overall difference between the dosage levels.

To investigate the interaction effect, we define the following population means:

μ_{11} = Population mean for Drug 1 and Dosage 10 mg,

μ_{12} = Population mean for Drug 1 and Dosage 20 mg,

μ_{21} = Population mean for Drug 2 and Dosage 10 mg, and

μ_{22} = Population mean for Drug 2 and Dosage 20 mg.

The null and alternative hypotheses for the interaction are

$$H_0: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22}$$

$$H_1: \mu_{11} - \mu_{21} \neq \mu_{12} - \mu_{22}.$$

One way to evaluate the interaction variability is to first calculate the variability among the four cell means. We call this the cells SS , defined as

$$\begin{aligned} SS_{CELLS} &= N_{11}(\bar{X}_{11} - \bar{X}_T)^2 + N_{12}(\bar{X}_{12} - \bar{X}_T)^2 \\ &\quad + N_{21}(\bar{X}_{21} - \bar{X}_T)^2 + N_{22}(\bar{X}_{22} - \bar{X}_T)^2 \\ SS_{CELLS} &= 3(7.0 - 5.5)^2 + 3(9.0 - 5.5)^2 \\ &\quad + 3(4.0 - 5.5)^2 + 3(2.0 - 5.5)^2 \\ SS_{CELLS} &= 3(1.5)^2 + 3(3.5)^2 + 3(-1.5)^2 + 3(-3.5)^2 \\ SS_{CELLS} &= 3[2.25 + 12.25 + 2.25 + 12.25] \\ SS_{CELLS} &= 3[29] \\ SS_{CELLS} &= 87. \end{aligned}$$

There are four cell means or four different treatments among the two drugs and two dosage levels. In a larger factorial design, we might have more than two levels of Factor A, more than two levels of Factor B, or more than two levels of both. If the total number of cells or treatment combinations is k , then the value of CELLS degrees of freedom would be

$$df_{CELLS} = k - 1.$$

In the present case we have

$$df_{CELLS} = 4 - 1 = 3.$$

The interaction sum of squares, SS_{AB} , can be found from

$$\begin{aligned} SS_{AB} &= SS_{CELLS} - SS_A - SS_B \\ SS_{AB} &= 87 - 75 - 0.0 \\ SS_{AB} &= 12. \end{aligned}$$

$$df_{AB} = df_A \times df_B = 1 \times 1 = 1.$$

From the above we have

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}} = \frac{12.0}{1} = 12.0.$$

We test the interaction null hypothesis with

$$F_{AB} = \frac{MS_{AB}}{MS_{WG}} = \frac{12.0}{1} = 12.0.$$

Testing at $\alpha = .01$, we reject H_0 because $12.0 > 11.26$. The greater adjustment for Drug 1 compared with Drug 2 is significantly greater for 20 mg ($7.0 = 9.0 - 2.0$) than for 10 mg ($3.0 = 7.0 - 4.0$).

Some researchers like to test simple main effects to aid in the interpretation of a significant interaction. For example, the difference between adjustment scores of the two drugs when both are administered at 20 mg is 7.0. It can be shown that the difference is significant at the .01 level. It can also be shown that the difference of 3.0 for 10 mg is also significant at the .01 level. Thus, these two simple main effects are not helpful for the data in Table 1 because the same pattern of significance is found for both simple main effects. The interpretation that

the 7.0 difference is significantly greater than the 3.0 difference can be made without any additional testing.

There are three assumptions for each of the F tests in a factorial design. They are the same three assumptions as for the independent-samples t test and the one-way, independent-groups ANOVA. They are as follows:

1. Independent observations are assumed in each cell of the design. This includes random assignment to groups. Failure of this assumption can completely invalidate the results of the F tests.
2. The populations are normally distributed. The F tests are relatively robust to failure of the normality assumption. With markedly nonnormal populations, alternative tests, such as nonparametric procedures, must be considered.
3. Population variances are assumed to be equal. With equal sample sizes that are not too small, the F tests are relatively robust to failure of the equal variances assumption. Equal N of 7 to 15 may be needed, depending on the degree of inequality of variances.

Although equal N s are not a requirement of factorial designs, they are highly recommended, even if the equal variances assumption is satisfied. Equal sizes are necessary to maintain independence of the three effects in factorial design. In fact, unequal cell sizes lead to a number of complications in the testing and interpretation of the effects of factorial design.

—Philip H. Ramsey

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Multivariate Analysis of Variance (MANOVA)

Further Reading

- Boik, R. J. (1979). Interactions, partial interactions, and interaction contrasts in the analysis of variance. *Psychological Bulletin*, 86, 1084–1089.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Maxwell, S. E., and Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical principles in experimental designs*. New York: McGraw-Hill.

FAGAN TEST OF INFANT INTELLIGENCE

The Fagan Test of Infant Intelligence (FTII, published by Infantest Corporation) purports to index infant intelligence by assessing infants' information processing capacities. The degree to which continuity exists between infant cognitive capacities and later intelligence has interested researchers since the 1930s. Until the early 1980s, intelligence in infancy was thought to be unrelated to intellectual functioning in childhood and adulthood. This sentiment was based on the lack of association between researchers' indices of infant intelligence, which assessed sensory-motor functioning, and adult measures of intelligence.

The FTII was developed in the early 1980s by Joseph F. Fagan, III, to assess infant information processing capacities such as visual recognition memory, habituation, and discrimination and relate them to intellectual functioning later in life. The FTII and other tests tapping infant information processing have led researchers to believe today that there is a relation between infant intellectual capacities and later intellectual functioning.

The FTII procedure is conducted on infants between 3 and 12 months of age and rests on the well-known tendency of infants to gaze more at novel stimuli than at familiar stimuli. The standard procedure involves two phases; the timing of each phase varies according to the age of the infant. In the "familiarization phase," the infant sitting in the parent's lap is presented with two identical stimuli, such as faces or geometric patterns, until the infant gazes at them for a predetermined amount of time. In the "test phase," one of the familiar stimuli is then paired with a novel stimulus. This procedure is repeated approximately 10 times during one sitting. The proportion of total looking time during which the infant gazes at the novel

stimulus is used to derive a novelty preference score thought to reflect some of the cognitive capacities of the infant.

There is mixed evidence as to whether the FTII predicts later intellectual functioning. A quantitative review of several empirical investigations suggests that the FTII and other similar measures of infant cognitive functioning positively correlate with later intellectual performance. There is also evidence that the FTII is a valid screening device in infancy that predicts mild to severe mental retardation with up to 80% accuracy. Nevertheless, the FTII has been criticized for its low predictive validity and lack of reliability. As a whole, the FTII assesses infant cognitive capacities, but some controversy surrounds the degree to which the procedure accurately predicts later intelligence.

—*Matthew J. Hertenstein and Lauren E. Auld*

See also Culture Fair Intelligence Test; Gf-Gc Theory of Intelligence; Intelligence Quotient; Intelligence Tests

Further Reading

Fagan, J. F., & Detterman, D. K. (1992). The Fagan Test of Infant Intelligence: A technical summary. *Journal of Applied Developmental Psychology*, 13, 173–193.

FTII description and other measures of infant cognitive assessment: <http://ehp.niehs.nih.gov/members/2003/6205/6205.html#comp>

FAMILY ENVIRONMENT SCALE

The Family Environment Scale (FES; published by Mind Garden, www.mindgarden.com) is composed of 10 subscales that measure the actual, preferred, and expected social environments of all types of families. The 10 subscales assess three sets of dimensions: (a) relationship dimensions (cohesion, expressiveness, conflict); (b) personal growth or goal orientation dimensions (independence, achievement orientation, intellectual-cultural orientation, active-recreational orientation, moral-religious emphasis); and (c) system maintenance dimensions (organization, control). The relationship and system maintenance dimensions

primarily reflect internal family functioning; the personal growth or goal orientation dimensions primarily reflect the linkages between the family and the larger social context.

The FES has three forms:

1. The Real Form (Form R) measures people's perceptions of their current family or their family of origin. This form is used to assess individuals' perceptions of their conjugal and nuclear families, formulate clinical case descriptions, monitor and promote improvement in families, focus on how families adapt to life transitions and crises, understand the impact of the family on children and adolescents, and predict and measure the outcome of treatment.

2. The Ideal Form (Form I) measures people's preferences about an ideal family environment. This form is used to measure family members' preferences about how a family should function; assess family members' value orientations and how they change over time, such as before and after family counseling; and identify areas in which people want to change their family.

3. The Expectations Form (Form E) measures people's expectations about family settings. This form is used in premarital counseling to clarify prospective partners' expectations of their family, help members of blended families describe how they expect their new family to function, and identify parents' expectations about their family after a major life transition, such as retirement or the youngest child's leaving home.

The FES manual presents normative data on 1,432 normal families and 788 distressed families, describes the derivation and application of a family incongruence score that assesses the extent of disagreement among family members, presents psychometric information on the reliability and stability of the subscales, and covers the research applications and validity of the subscales. The manual includes a conceptual model of the determinants and outcomes of the family environment and reviews studies focusing on families of youth with behavioral, emotional, or developmental

disabilities; families with a physically ill child; families with a history of physical or sexual abuse; and families of patients with medical and psychiatric disorders. The FES has also been used to focus on the relationship between the family environment and child development and adult adaptation and on families coping with life transitions and crises, such as parent or child death, unemployment and economic deprivation, immigration and acculturation, and combat and war.

—Rudolf H. Moos and Bernice S. Moos

See also Social Climate Scale

Further Reading

Moos, R. (2001). *The Family Environment Scale: An annotated bibliography* (3rd ed.). Redwood City, CA: Mind Garden.

Moos, R., & Moos, B. (1994). *Family Environment Scale manual* (3rd ed.). Redwood City, CA: Mind Garden.

FILE DRAWER PROBLEM

A meta-analysis is a quantitative summary or synthesis of findings of studies that focus on a common question; one example is a quantitative synthesis of results of studies that focus on the efficacy of psychotherapy. Unfortunately, studies that are included in a meta-analysis can be unrepresentative of all the methodologically sound studies that address this common question, so the combined results of the studies in the meta-analysis can be misleading. Included studies may be unrepresentative because of the well-documented “publication bias,” which refers to a bias against publication of results of studies that do not yield statistically significant results. Because of this bias, results of studies that are not statistically significant often (a) do not appear in print (either as journal articles or as published abstracts of presentations), (b) wind up tucked away in researchers’ “file drawers,” and (c) remain undetected or inaccessible to meta-analysts. In the most extreme case of the “file drawer problem,” the collection of studies included in

a meta-analysis consists exclusively of those that yielded results significant at the conventional .05 level.

The most popular method of dealing with the file drawer problem involves calculation of Robert Rosenthal’s Fail-Safe-N (FSN). The FSN—which was derived under the (questionable) assumptions that (a) the studies targeted by meta-analyses use two-tailed (nondirectional) tests and (b) the studies in the file drawers average null results—is an estimate of the minimum number of unpublished studies (tucked away in file drawers) that would threaten the validity of significant combined results of a meta-analysis. For example, for a well-known 1982 meta-analysis (by Landman and Dawes) focusing on a set of 42 studies of efficacy of psychotherapy that were considered (by the meta-analysts) to be appropriately controlled, the FSN was 461. Since the combined results of this meta-analysis indicated statistically significant beneficial effects of psychotherapy, it was inferred (by FSN users) that there would have to exist at least 461 unpublished file drawer studies (averaging null results) to threaten the validity of this conclusion. Although there are no firm guidelines for interpretation of FSNs, Rosenthal suggested using $FSN_c = K(5) + 10$ as a critical value or rule of thumb (where K = number of studies in the meta-analysis); thus only FSNs below FSN_c would be considered to threaten significant combined results of a meta-analysis. For the Landman and Dawes meta-analysis, the FSN of 461 is well above the FSN_c of 220 (i.e., $42(5) + 10 = 220$), suggesting to users of the FSN that the file drawer problem was negligible in this meta-analysis.

—Louis M. Hsu

See also Meta-Analysis

Further Reading

Darlington, R. B., & Hayes, A. F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. *Psychological Methods*, 4, 496–515.

Hsu, L. M. (2000). Effects of directionality of significance tests on the bias of accessible effect sizes. *Psychological Methods*, 5, 333–342.

Hsu, L. M. (2002). Fail-Safe Ns for one- and two-tailed tests lead to different conclusions about publication bias. *Understanding Statistics*, 1, 85–100.

- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109–135.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

FISHER, RONALD AYLMEYER (1890–1962)

Ronald Aylmer Fisher was a statistician, eugenicist, evolutionary biologist, and geneticist who helped lay the foundations of modern statistical science. At an early age, Fisher's special abilities were apparent. Because of his poor eyesight—extreme myopia—he was forbidden to read with the aid of electric lights. In the evenings, his mother and teachers would read to him and provide instruction without visual aids. Fisher's exceptional ability to solve mathematical problems in his head and his geometrical approach to statistical problems are attributed to this early form of instruction.

Fisher obtained a bachelor's degree in mathematics from the University of Cambridge in 1912. The following year, he took a statistical job with the Mercantile and General Investment Company of London, a position he held for two years. Because of his poor eyesight, he was rejected for military service during World War I and spent the war years teaching mathematics and physics at various public schools. In 1917, he married Ruth Eileen Guinness. The marriage produced two sons and eight daughters.

In 1919, Fisher accepted the newly created position of statistician at Rothamsted Experimental Station, approximately 20 miles north of London. The career move proved to be a fortunate choice. It brought him into close contact with researchers who were concerned with the interpretation of agricultural field experiments and with laboratory and greenhouse experiments. Rothamsted provided an environment in which Fisher was free to pursue his interests in genetics and evolution; carry out breeding experiments on mice, snails, and poultry; and develop statistical methods for small samples.

While at Rothamsted, Fisher invented the analysis of variance and revolutionized the design of experiments. The publication of Fisher's books *Statistical Methods for Research Workers* in 1925 and *The Design of Experiments* in 1939 gradually led to the acceptance of what today are considered the cornerstones of good experimental design: randomization, replication, local control or blocking, confounding, randomized blocks, and factorial arrangements. Other notable contributions include the concept of likelihood and the maximum likelihood estimator, the development of methods suitable for small samples, the discovery of the exact distribution of numerous statistics derived from small samples, the Fisher information measure, and contributions to hypothesis testing. In 1933, Fisher succeeded Karl Pearson as the Galton Professor of Eugenics at University College, London. Ten years later, he accepted an appointment as the Arthur Balfour Chair of Genetics at Cambridge, a position he held until his retirement in 1957.

Friends described Fisher as charming and warm but possessing a quick temper and a devotion to scientific truth as he saw it. The latter traits help explain his long-running disputes with Karl Pearson and Pearson's son Egon. Fisher was the recipient of numerous honors and was created a Knight Bachelor by Queen Elizabeth in 1952.

—Roger E. Kirk

See also Analysis of Variance (ANOVA)

Further Reading

Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley.

Ronald A. Fisher biographical essay: http://en.wikipedia.org/wiki/Ronald_Fisher

FISHER EXACT PROBABILITY TEST

The Fisher exact probability test (also called the Fisher-Irwin test) is one of several tests that can be

used to detect whether one dichotomous variable is related to another. The rationale of this test, as well as its principal advantages and limitations, can be presented in the context of the following hypothetical small randomized experiment designed to determine whether a dichotomous “treatment” variable (Drug vs. Placebo) is related to a dichotomous “outcome” variable (Survival vs. Death).

A physician believes that a new antiviral drug might be effective in the treatment of SARS (severe acute respiratory syndrome). Assume that the physician carries out a randomized double-blind drug efficacy study involving 6 SARS patients (designated A, B, C, D, E, and F), 3 of whom (say, A, B, and C) were randomly selected from this group and given the drug and the remaining 3 of whom (D, E, and F) were given a placebo. Four months later, the 3 patients who received the drug were still alive whereas the 3 patients who received the placebo were not.

Results of this study may be summarized in a 2×2 table (see Table 1, which has 2 rows and 2 columns, ignoring row and column totals). More generally, results of any randomized treatment efficacy study involving dichotomous treatment and outcome variables may be summarized using the notation shown in Table 2. Do the results in Table 1 support the belief that the new drug is effective (relative to the placebo)? Or, more generally, do results of a randomized study that can be summarized as in Table 2 support the belief that the two dichotomous variables are related—for example, that patient outcomes are related to the treatments to which they have been exposed?

The fact that, in the physician’s study, all the drug patients survived and all the placebo patients died (Table 1) would seem consistent with the belief that the patient outcomes were related to treatments they received. But is there a nonnegligible probability that such a positive result could have occurred if the treatment had in fact been unrelated to the outcome? Consistent with absence of relation of outcomes to treatments, let us hypothesize (this will be called the null hypothesis, and designated H_0 hereafter) that patients A, B, and C, who actually survived, would have survived whether they received the drug or the

Table 1 Results of a 2×2 Design

	<i>Drug</i>	<i>Placebo</i>	<i>Row Totals</i>
Survived	$X = 3$	0	3
Died	0	3	3
Column total	3	3	6

Table 2 Summary of Study Results

	<i>Treatment 1</i>	<i>Treatment 2</i>	<i>Row Totals</i>
Success	$X = a$	b	N_s
Failure	c	d	N_f
	N_1	N_2	N

placebo, and that D, E, and F, who actually died, would have died whether they received the drug or the placebo. Now we ask, Would the positive results in Table 1 have been unlikely if this H_0 had been true?

This H_0 has two important implications that are relevant to answering the question: First, the total number of survivors and nonsurvivors would be 3 and 3, respectively, regardless of which 3 patients were selected to receive the drug (and which other 3 received the placebo), so that the marginal totals of Table 1 would be fixed regardless of results of the randomization. Second, the number of survivors among the drug patients (which will be designated X hereafter), as well as the other 3 entries in the 2×2 table, would be determined by which patients were selected to receive the drug. For example, with the selection A, B, and C, the number of drug patients who would survive (X) would be 3, and the results of the study would be as displayed in Table 1. But with the selection of A, B, and D, then X would be 2, and the four cells of the 2×2 table would then have entries of [2 1] for row 1 and [1 2] for row 2 (note that since marginal totals of Table 1 are fixed, irrespective of which 3 patients are selected to receive the drug, knowledge of X determines entries in the other three cells of the 2×2 table; these entries are therefore redundant with X). With fixed marginal totals, the variable X is clearly relevant to tenability of the H_0 relative to the hypothesis that

Table 3 Patient Selection Possibilities for Receiving Drug

ABC(3)								
ABD(2)	ABE(2)	ABF(2)	ACD(2)	ACE(2)	ACF(2)	BCD(2)	BCE(2)	BCF(2)
ADE(1)	ADF(1)	AEF(1)	BDE(1)	BDF(1)	BEF(1)	CDE(1)	CDF(1)	CEF(1)
DEF(0)								

the drug is effective. There were, in fact, 20 possible ways in which the 3 patients who were to receive the drug could have been selected; these 20 selections are listed in Table 3, together with values of X they would have determined under the H_0 .

Since the selection was random (meaning that all 20 possible selections were equiprobable), it is apparent that the probabilities of $X = 0, 1, 2$, and 3 given H_0 (designated $P(X | H_0)$) can be determined by counting how many of the 20 assignments yield each value of X and dividing this count by 20. Table 4 lists the probabilities obtained by applying this enumeration method.

Table 4 Probabilities Obtained by Applying This Enumeration Method

X	0	1	2	3
$P(X H_0)$	1/20 = .05	9/20 = .45	9/20 = .45	1/20 = .05

Table 4 shows that the very positive result of the physician's study (in particular, $X = 3$) is improbable under the H_0 that the outcome (survival, death) was unrelated to treatment assignment for each of the 6 patients (in particular, $P(X = 3 | H_0) = 0.05$). This result, however, would be consistent with the hypothesis that the new drug worked. Table 4 also implies that a researcher who decided to reject the H_0 (i.e., to conclude that outcomes were related to treatments) if $X = 3$ would be taking a risk of 0.05 of making a Type I error (defined as "rejection of a true null hypothesis"). Note also that if a researcher decided to reject the H_0 if $X \geq 2$, then the researcher would be taking a very large risk (viz., $.45 + .05 = .50$) of making a Type I error.

Determining Fisher Exact Probabilities From Hypergeometric Distributions

In general (see Table 2), given (a) random assignment of N_1 participants (drawn from a pool of N participants) to Treatment 1 and $N_2 = (N - N_1)$ participants to Treatment 2, (b) N_s were observed to "Succeed" (e.g., survive) and N_f were observed to "Fail" (e.g., die), and (c) the null hypothesis (H_0) that the outcome is unrelated to treatment assignment, the probability of ($X = a$) successes among participants exposed to Treatment 1 can be obtained using the formula

$$P(X = a | H_0) = \frac{\frac{N_s!N_f!}{a!(N_s - a)!(N_1 - a)!(N_f - N_1 + a)!}}{\frac{N!}{N_1!(N - N_1)!}},$$

where any number (say, M) followed by an exclamation point ($M!$) is defined as follows:

$$M! = M(M-1)(M-2) \dots (3)(2)(1) \\ \text{(e.g., } 6! = (6)(5)(4)(3)(2)(1) = 720\text{),}$$

and where $0! = 1! = 1$. For example, the $P(X = 2 | H_0)$ for Table 1 would be

$$P(X = 2 | H_0) = \frac{\frac{3!}{2!(3-2)!} \frac{3!}{1!(3-1)!}}{\frac{6!}{3!(6-3)!}} \\ = \frac{\frac{(3)(2)(1)}{(2)(1)(1)} \frac{(3)(2)(1)}{(1)(2)(1)}}{\frac{(6)(5)(4)(3)(2)(1)}{(3)(2)(1)(3)(2)(1)}} = \frac{9}{20} = .45$$

Note that the numerator yields the number of equiprobable selections of N_1 from N participants that would result in $X = 2$ (cf. the second line in Table 3) and that the denominator (obtained from $N! / (N_1!(N - N_1)!)$) is a count of the total number of possible equiprobable selections of N_1 from N (cf. all four lines in Table 3). The above formula, which calculates $P(X | H_0)$ using efficient counting principles, is called the general term of the hypergeometric distribution, and the probability distribution of X that can be used to generate (e.g., Table 1) is called the hypergeometric distribution.

The Fisher exact test is therefore appropriate for randomized designs (with random selection of N_1 participants, drawn from a pool of N participants, for assignment to Treatment 1 and $N_2 = N - N_1$ participants for assignment to Treatment 2, and with N_s Successes and N_f Failures) to address questions about relations of dichotomous “treatment” and “outcome” variables (see Table 2). The null hypothesis (H_0) in these designs is that, for each of the N participants, the outcome is unrelated to the treatment assignment. This H_0 implies that the selection (treatment assignment) does not affect the total numbers of Successes or Failures observed in the study (i.e., fix the row totals, N_s and N_f) but determines a , the value of X (see Table 2), as well as values of b , c , and d (since the value of a determines the values of b , c , and d , given the fixed marginal totals). Given randomization, X will have the hypergeometric distribution under H_0 . Evidence against H_0 , in a given study, consists of obtained values of X (the number of Treatment 1 patients who Succeed) so extreme as to have been very unlikely to have occurred had H_0 been true (say $p \leq .05$, as determined from the hypergeometric distribution). When the researcher’s prediction or experimental hypothesis implies high values of X , the expression “exact probability” in the Fisher “exact probability” test usually refers to $P(X \geq X_{\text{obtained}} | H_0)$, where X_{obtained} is the value of X obtained in the study and where the probabilities that are cumulated to calculate $P(X \geq X_{\text{obtained}} | H_0)$ are determined by the right side or “right tail” of the hypergeometric distribution. For example (see Table 1), if in the physician’s study, $X_{\text{obtained}} = 2$ (i.e., 2 of the 3 patients who were given the

drug survived), then $P(X \geq X_{\text{obtained}} | H_0)$ would be $.45 + .05 = .50$. This probability is called a one-tailed p value of the Fisher exact test. Similarly, when the experimental hypothesis implies low values of X , the “exact probability” of Fisher’s test results in another one-tailed p value, defined as $P(X \leq X_{\text{obtained}} | H_0)$, the sum of probabilities of $X \leq X_{\text{obtained}}$, located on the left side or left tail of the hypergeometric distribution. A “two-tailed p value” for the Fisher test would be the sum of probabilities of X at least as extreme as X_{obtained} in both tails of the hypergeometric distribution. For example, if $X_{\text{obtained}} = 3$, results at least as extreme in the hypergeometric distribution would be $X = 3$ and $X = 0$, and the two-tailed p value of the Fisher test would therefore be $.05 + .05 = .10$ (see Table 4).

Assumptions and Limitations

The above information draws attention to two important characteristics of the Fisher exact test. First, the exact probability generated for this test is based on two important assumptions: (a) that both column and row marginal totals of the 2×2 table are fixed under the H_0 and (b) that the $[N! / (N_1!(N - N_1)!)]$ possible results of the experiment are all equiprobable under the H_0 . These assumptions are required for the derivation of the hypergeometric distribution that is used to calculate the “exact probabilities” of this test. In the randomized design illustrations (above), assumption (a) is satisfied by the researcher’s decision to select N_1 (drawn from the pool of N) participants for assignment to Treatment 1 and $N_2 = N - N_1$ participants for assignment to Treatment 2 (this decision fixes the column totals) and by the definition of the H_0 (which fixes the row totals, as explained above). Assumption (b) is satisfied, in these illustrations, by the random assignment. Second, the exact probability associated with the Fisher test is relevant, in these illustrations, to inferences about the relation of the outcome to the treatment *only for the N participants included in the study*. For example, the p value of .05 corresponding to the results of the physician’s study warrants the conclusion (assuming that a risk of .05 of a Type I error is acceptable) that the H_0 can be rejected *for the 6 patients included in the physician’s study*. The

p value of the Fisher test does not support statistical inferences about efficacy of the drug with patients other than the 6 who took part in the study (although nonstatistical inferences may be possible).

The Fisher exact probability test can be applied to 2×2 table data of studies that do not involve random assignment of participants to conditions, as long as the hypergeometric distribution assumptions (viz., fixed marginal totals for both rows and columns of the 2×2 table and equiprobability under H_0 of the $[N! / (N_1!(N - N_1)!)]$ are plausible under the researcher's H_0). It is noteworthy that the experiment R. A. Fisher originally chose to illustrate his exact test did not involve random assignment of N participants to treatments but instead involved a selection made by a single participant. Fisher's experiment was designed to test a tea-drinking lady's claim that she could taste the difference between two cups of tea with milk, one in which the milk was added prior to the tea, and the other in which the tea was added prior to the milk. Fisher's experiment consisted of presenting the lady with eight cups (in random order) after informing her that in half of the cups, milk had been added first and in the other half, tea had been added first, and asking her to select the four in which the milk had been added first. The H_0 was that her claim was false and that her selection of four from the eight cups could then in effect be viewed as a random selection of four from eight cups. The number of successes (X = number of cups in which milk was added first) for the four cups she selected, which could be 0, 1, 2, 3, or 4, is relevant to testing her claim and can be viewed as a hypergeometric random variable if the H_0 is true. Clearly, the larger the value of X , the stronger the evidence against the H_0 (and indirectly for the lady's claim). Assuming that the lady obtained $X = 4$ successes, the exact one-tailed p level of the Fisher exact probability test would be $1/70 = .014$ (which may be verified by enumeration or by applying the general term of the hypergeometric distribution); in other words, it is improbable that she could have had that many successes if her claim had been false (and if her selection had in effect been a random selection of four from the eight cups). (Historical note: R. A. Fisher, who designed this interesting little experiment, once told M. G. Kendall that he never actually carried it out.)

A researcher who considers a p level of .014 too large for rejection of the H_0 could easily modify the study by increasing the number of cups so that more convincing evidence against the H_0 could be obtained. Thus, for example, if from a set of 16 cups the lady correctly identified the 8 cups in which milk had been poured first, the one-tailed p level of the Fisher exact test would be $1/12,870 = .0000777$ (which can easily be determined from the general term of the hypergeometric distribution). This result would provide very convincing evidence against the hypothesis (H_0) that her selection had been a random selection. However, as in the physician's study (above), the Fisher test would not allow statistical inferences about persons not included in the study: In particular, a very small p value for the tea-tasting experiment would not allow statistical inferences about tea-tasting abilities of any person other than the lady who took part in the study.

Using the Computer

SYSTAT offers two-tailed p values for the Fisher exact probability test as an option in its "cross-tab" module when the cross-tabulation involves two dichotomous variables. The SYSTAT output for the 16-cup tea-tasting experiment in which the lady correctly selected all (i.e., $X = 8$) of the cups in which milk had been poured first includes the 2×2 table and the exact two-tailed p value of the Fisher test:

	Milk First	Tea First	Total
Selected	8	0	8
Not selected	0	8	8
Total	8	8	16

Fisher exact test (two-tailed) probability = .000155.

—Louis M. Hsu

Further Reading

- Fisher, R. A. (1971). *The design of experiments*. New York: Hafner. (Original work published in 1935)
- Hodges, J. L., Jr., & Lehmann, E. L. (1970). *Basic concepts of probability and statistics* (2nd ed.). San Francisco: Holden-Day.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Erlbaum.

Fisher's exact test calculation: <http://www.unc.edu/~preacher/fisher/fisher.htm>

FISHER-IRWIN TEST

See FISHER EXACT PROBABILITY TEST

FISHER'S LSD

The analysis of variance (ANOVA) can be used to test the significance of the difference between two or more means. A significant overall F test leads to the rejection of the full null hypothesis that all population means are identical. However, if there are more than two means, then some population means might be equal. R. A. Fisher proposed following a significant overall F test with the testing of each pair of means with a t test applied at the same level α as the overall F test. No additional testing is done following a non-significant F because the full null hypothesis is not rejected. This procedure was designated the least significant difference (LSD) procedure.

If there are exactly three means in the ANOVA, the probability of one or more Type I errors is limited to the level α of the test. However, with four or more means, that probability can exceed α . A. J. Hayter proposed a modification to LSD that limits the probability of a Type I error to α regardless of the number of means being tested.

If the number of means is k then Tukey's honestly significant difference (HSD) procedure can be used to test each pair of means in an ANOVA using critical values from the Studentized range distribution. Hayter proposed replacing Fisher's t tests with the HSD critical values that would be used with $k - 1$ means even though the number of means is k . That is, a significant F test is followed by testing all pairs of means from the k means with the HSD critical value for $k - 1$ means.

Illustrative Example

Table 1 presents a hypothetical data set in which four groups containing five observations each produce a within-groups MS of 2.0. An independent-groups ANOVA applied to such data would produce an overall $F = 22.62$, which would exceed the critical value, $F_{.95}(1,16) = 3.24$.

Table 1 A Hypothetical Data Set in Which Four Groups Contain Five Observations

Group 1	Group 2	Group 3	Group 4
2.00	4.31	6.61	9.00

Following the significant F test, the LSD procedure requires testing all pairs of means. In Hayter's modification, a critical difference for all pairs at the .05 level is obtained from the formula

$$CD = SR_{.95, k-1, v} \sqrt{\frac{MS_E}{N}},$$

where

$SR_{.95, k-1, v}$ is the Studentized range statistic,

k is the number of means,

v is the error degrees of freedom,

MS_E is the error term (in this case the Mean Square within groups), and

N is the common group size.

For the data in Table 1, we obtain the result

$$\begin{aligned} CD &= 3.65 \sqrt{\frac{2.0}{5}} = 3.65 \sqrt{0.4} \\ &= 3.65(.63245) = 2.31. \end{aligned}$$

Table 2 presents the six pairwise differences and shows that all pairs are significantly different except for Group 2 and Group 3. That difference of 2.30 is less than the critical difference of 2.31.

Table 2 The Six Possible Pairwise Differences

	Group 1 2.0	Group 2 4.31	Group 3 6.61	Group 4 9.00	Critical Difference
2.0	–	2.31*	4.61*	7.00*	2.31
4.31		–	2.30	4.69*	
6.61			–	2.39*	

* = Significantly different at $\alpha = .05$.

The Hayter-Fisher version of LSD will always be more powerful than Tukey's HSD for testing all pairwise differences following an ANOVA F test. In the case of testing exactly three means, the Hayter-Fisher version gives the same results as the original LSD.

Most computer packages apply the original version of LSD. In the case of testing exactly three means, those results will be accurate. However, when testing more than three means, such packages will risk excessively high Type I error rates.

Most computer packages also provide Tukey's HSD procedure. Following a significant overall F test, any pair of means found significantly different by HSD will also be significantly different by the Hayter-Fisher version of LSD. Also, any pair of means found not to be significantly different by the original LSD will also not be significantly different by the Hayter-Fisher version. Thus, in most cases the results of the Hayter-Fisher procedure can be found from a computer package that provides HSD and the original LSD. In cases where a pair of means is not significant by HSD but is significant by the original LSD, the final decision for the Hayter-Fisher can be determined quite easily as described above. Of course, computer packages could easily modify their procedures to offer the Hayter-Fisher method.

—Philip H. Ramsey

See also Bonferroni Test; Post Hoc Comparisons

Further Reading

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81, 1000–1004.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods*, 7, 504–523.

FISHER'S Z TRANSFORMATION

Fisher's Z transformation is a procedure that rescales the product-moment correlation coefficient into an interval scale that is not bounded by ± 1.00 . It may be used to test a null hypothesis that an obtained correlation is significantly different from some hypothesized value (usually a nonzero value, because a t test is available to test whether $\rho = 0$), to test the significance of the difference between two independent correlations, to find the average of several correlations, or to form a confidence interval (CI) for a correlation coefficient.

Like all statistics, the correlation coefficient is subject to sampling variation. For a given population, the sample correlation coefficient (r) has a sampling distribution around its population parameter, ρ (Greek lowercase letter rho), and this distribution has a standard error, the standard error of the correlation coefficient, σ_r . However, the sampling distribution of r has a different shape, depending on the value of ρ . The possible values of r are limited to the range between $+1.0$ and -1.0 . If the value of ρ is about zero, sample deviations can occur equally in either direction, and the distribution is symmetric. However, as ρ departs from zero, this symmetry is lost because one end of the sampling distribution is more restricted than the other due to the limiting values of ± 1.0 . There is more space on one side of the parameter than on the other. For extreme values of ρ , the sampling distribution of r becomes markedly skewed. The skew becomes noticeable at about $\rho = .40$.

A solution to this problem was developed by R. A. Fisher, who proposed a transformation for r that would normalize the distribution. The resulting index is called Fisher's Z or Z_F . Note that this is not the same quantity as the standard score or the critical ratio, each of which is a deviation of an observed value from the

distribution mean, divided by the standard deviation. The statistic Fisher developed,

$$Z_F = 1/2 \ln \left(\frac{1+r}{1-r} \right), \quad (1)$$

has the advantage that its sampling distribution is almost exactly normal for any value of ρ and has a standard error of

$$\sigma_{Z_F} = \frac{1}{\sqrt{N-3}}. \quad (2)$$

This standard error does not depend on the value of ρ , unlike the standard error used in the t test for testing the hypothesis that $\rho = 0$. Also, σ_{Z_F} is calculated, not estimated, so the test statistic is a critical ratio Z and is more powerful than the t test.

Because r can have only a limited number of values (to 2 decimal places) and the distribution is symmetric around zero, some statistics books contain tables for transforming r to Z_F and Z_F back to r . However, the equations for transforming a correlation to the metric of Z_F and back again are so simple that they are easily entered into a cell of a spreadsheet or programmed as a macro. The equation for the reverse transformation is

$$r = (\exp(2 * Z) - 1) / (\exp(2 * Z) + 1) \quad (3)$$

$$= \frac{e^{2Z} - 1}{e^{2Z} + 1},$$

where the first form of the expression is in the notation used for computations by EXCEL and the second is in standard notation. However, many spreadsheets also include functions for converting to and from Z_F in their function libraries. For example, EXCEL has a function called FISHER for the r -to- Z_F transformation in Equation 1 and one called FISHERINV for the reverse transformation in Equation 3. Pasting FISHER into a cell produces the dialog box in Figure 1. Entering the observed r for X produces Z_F . In this example, $r = .80$ produces $Z_F = 1.0986$.

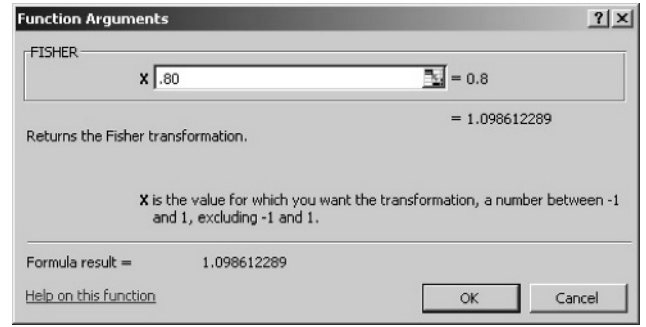


Figure 1 Excel Screen Image for Transforming r to Z_F

The process is reversed by pasting FISHERINV into a cell. In the example in Figure 2, we have reversed the transformation to recover our original correlation of .80.

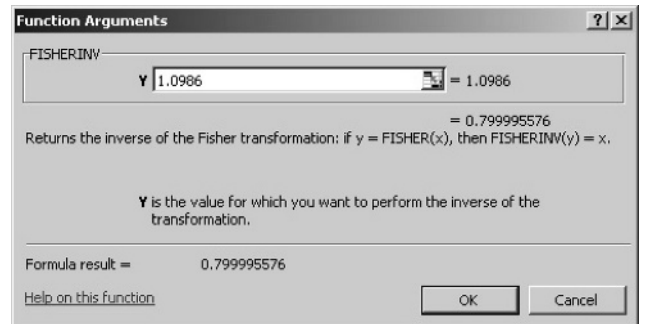


Figure 2 Excel Screen Image for Transforming Z_F to r

Applying Z_F to Hypothesis Testing

Sometimes we might wish to test a hypothesis that ρ is some value other than 0. Suppose we wish to test the null hypothesis that the correlation between scores on a reaction-time measure (RT) and scores on Thorndike's Precarious Prognosticator of Potential Proficiency (the TP⁴) is $-.75$. Because the sampling distribution of r around a population parameter as high as this is definitely not symmetric, we must use Fisher's Z transformation. We transform the r and the ρ to Z_F s and carry out the test using the formula

$$Z = \frac{Z_{F_r} - Z_{F_\rho}}{\frac{1}{\sqrt{N-3}}}, \quad (4)$$

where Z_F is the transformed value of the observed correlation and Z_{F_p} is the transformed value of the hypothesized population correlation. If Z from Equation 4 exceeds the usual critical value (1.96, 2.58, or another value we might choose), we would reject the null hypothesis that $\rho = 0.75$.

Suppose we have data from a sample of 68 participants, and the observed correlation between RT and the TP⁴ is $-.59$. The transformed values are $Z_{F_r} = -.678$ and $Z_{F_p} = -.973$, so with this sample size, we obtain

$$Z = \frac{-.678 - (-.973)}{\sqrt{\frac{1}{65}}} = \frac{+.295}{.124} = +2.38.$$

Since a Z of 2.38 exceeds the critical value of 1.96, we would reject the hypothesis that $\rho = -.75$ in the population that produced this sample at the $p \leq .05$ level. That is, it is unlikely that this sample was drawn from a population in which $\rho = -.75$. The fact that the Z is positive suggests that the sample was probably drawn from a population in which ρ is less than $-.75$ (that is, not as large a negative correlation).

We can see how important Fisher's Z transformation is if we compare the result above with what we would get if we tested the same hypothesis without the transformation. The appropriate equation (see t test of a correlation coefficient) produces

$$t_{66} = \frac{-.59 - (-.75)}{\sqrt{\frac{(1 - .55^2)}{(68 - 2)}}} = \frac{.16}{.099} = 1.62.$$

A t of 1.62 with $df = 66$ is not significant, so we would not reject the hypothesis that $\rho = -.75$. Had we been testing the hypothesis that $\rho = 0$, either test would have been appropriate, and either would have led to the same conclusion. The difference is caused by the asymmetry of the sampling distribution of r for large values of ρ .

Confidence Intervals for Correlations

A second situation that we occasionally encounter and that calls for Z_F is the need to place a CI for ρ on an observed r . For example, in the section above, we were able to conclude that it is unlikely that the population from which we drew our sample is one in which the correlation is $-.75$, but what is a reasonable range for this correlation? A CI provides an answer.

The first step is to determine Z_F for the obtained r . Next, calculate σ_{Z_F} using Equation 2. Using the appropriate normal deviate Z s for our chosen confidence level (Z_α) (for example, ± 1.96 for the 95% CI), we compute the upper and lower CI limits for Z_F using the usual linear transformations

$$CI_L = (-Z_\alpha)\sigma_{Z_F} + Z_{F_r}$$

and

$$CI_U = (+Z_\alpha)\sigma_{Z_F} + Z_{F_r}.$$

These limits, which are in the metric of Z_F , are then transformed into correlation values.

To find the 95% CI for the relationship between RT and the TP⁴, we would proceed as follows: The obtained r was $-.59$ computed on 68 participants, the Z_F was $-.678$, and $\sigma_{Z_F} = 0.124$.

The upper- and lower-limit values for the 95% CI in the metric of Z_F are

$$(-1.96)(0.124) + (-.678) = -.921$$

and

$$(+1.96)(0.124) + (-.678) = -.435.$$

This CI is symmetric in terms of probability and in terms of Z_F but not in terms of r . With FISHERINV, we can transform the values of Z_F into the metric of r . The CI in the scale of r goes from $-.409$ to $-.726$, with its center at $-.59$, reflecting the fact that the sampling distribution of r in this region is positively skewed. (The sampling distribution would be negatively skewed for positive correlations.)

Hypotheses About Two Correlations From Independent Samples

We occasionally encounter a situation where we want to test a null hypothesis that two correlations are equal or that they differ by a specified amount. If the two correlations involve the same variables in two different samples, such as testing whether a correlation is the same for men as it is for women, we can use Z_F . First, transform both r s to Z_F s, then calculate the standard error of the difference between two Z_F s:

$$\sigma_{(Z_{F_1}-Z_{F_2})} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}. \quad (5)$$

A critical ratio Z of the usual form then provides the test statistic for the null hypothesis that $\rho_1 = \rho_2$ ($\rho_1 - \rho_2 = 0$):

$$Z = \frac{Z_{F_{r_1}} - Z_{F_{r_2}} - (Z_{F_{\rho_1}} - Z_{F_{\rho_2}})}{\sigma_{(Z_{F_1}-Z_{F_2})}}. \quad (6)$$

Suppose we found the correlations between midterm and final examination scores for students in a statistics class. For the 17 men in the class, the correlation is 0.78, and for the 24 women, r is 0.55. The null hypothesis that $\rho_M = \rho_F$ is tested by finding

$$\sigma_{(Z_{F_1}-Z_{F_2})} = \sqrt{\frac{1}{17-3} + \frac{1}{24-3}} = \sqrt{.119} = .345$$

and

$$Z = \frac{(1.045 - .618) - 0}{.345} = \frac{+.427}{.345} = 1.24.$$

Clearly, we cannot reject the null hypothesis of equal correlations on the basis of these data, but we also cannot reach the conclusion that they are equal.

Averaging Correlations

Another application is the problem of averaging the correlation between two variables across several groups of people. For example, research on the Wechsler Intelligence Scale for Children sometimes involves several groups of children who differ in age. For some research questions, it is desirable to collapse the results across ages to obtain larger samples. It is not appropriate simply to consider all children to be a single sample, because cognitive ability is related to age. Therefore, correlations are computed within each age group and then averaged across ages.

Because the sampling distribution of the correlation coefficient is highly skewed for large values of ρ , the meaning of differences between correlations changes depending on where we are in the range of possible values. For this reason, we cannot use simple averaging techniques with r s. We can, however, use Z_F because it represents a variable on an interval scale. Whenever it is necessary to find an average of several correlations, particularly when they differ by more than about .10, the appropriate procedure is as follows:

1. Transform all r s to Z_F s.
2. Find $\bar{Z}_F$ (the mean of the Z_F s).
3. Then reverse the transformation using $\bar{Z}_F$ to find the mean r , or $\bar{r}$.

When the correlations being averaged are based on samples of different sizes, it is necessary to compute a weighted average. Under these conditions, the proper weighting is given by the formula in Equation 7:

$$\bar{Z}_F = \frac{(N_1 - 3)Z_{F_1} + (N_2 - 3)Z_{F_2} + \cdots + (N_K - 3)Z_{F_K}}{(N_1 - 3) + (N_2 - 3) + \cdots + (N_K - 3)}. \quad (7)$$

The mean correlation, $\bar{r}$, is found by transforming $\bar{Z}_F$ back into the metric of r using Equation 3. Hypotheses concerning this mean correlation should be tested using $\bar{Z}_F$, which has a standard error of

$$\sigma_{\bar{Z}_F} = \frac{1}{\sqrt{(N_1 - 3) + (N_2 - 3) + \cdots + (N_K - 3)}}. \quad (8)$$

The critical ratio test statistic for testing null hypotheses concerning $\bar{r}$ is

$$Z = \frac{\bar{Z}_F - Z_{F\rho}}{\sigma_{\bar{Z}_F}}. \quad (9)$$

As usual, $Z_{F\rho}$ is the Z_F of ρ , the correlation under the null hypothesis, which is usually zero but may take on any appropriate value.

Imagine that we have the correlation between the midterm exam scores and final exam scores for each of three classes in statistics. The data are

$r_1 = 0.77$	$N_1 = 12$	$Z_{F_1} = 1.020$
$r_2 = 0.47$	$N_2 = 18$	$Z_{F_2} = 0.510$
$r_3 = 0.25$	$N_3 = 57$	$Z_{F_3} = 0.255$

We may wish to find $\bar{r}$ and test the hypothesis that the mean correlation is zero. This is done by using Equations 7–9 to find

$$\begin{aligned} \bar{Z}_F &= \frac{9(1.020) + 15(.510) + 54(.255)}{9 + 15 + 54} \\ &= \frac{30.6}{78} = .392 \end{aligned}$$

$$\bar{r} = 0.373$$

$$\sigma_{\bar{Z}_F} = \frac{1}{\sqrt{78}} = .113$$

$$Z = .373/.113 = 3.30,$$

from which we can conclude that the mean correlation of 0.373 differs significantly from zero.

We can place a CI on this mean correlation in the usual way. For the 95% CI, we start with

$$\bar{Z}_F = .392 \quad \sigma_{\bar{Z}_F} = .113 \quad Z_{0.95} = \pm 1.96.$$

Finding the lower and upper limits of the CI for $\bar{Z}_F$ yields

$$\begin{aligned} CI_L &= -1.96(.113) + .392 = .171 \\ CI_U &= +1.96(.113) + .392 = .613. \end{aligned}$$

Converting these Z_F s back into correlations produces a CI centering on +.37 and running from +.17 to +.57.

—Robert M. Thorndike

See also Correlation Coefficient

Further Reading

- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.
- Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: Freeman.
- Thorndike, R. M. (1994). Correlational procedures in data analysis. In T. Husen & N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 1107–1117). New York: Pergamon.
- Thorndike, R. M., & Dinnel, D. L. (2001). *Basic statistics for the behavioral sciences*. Upper Saddle River, NJ: Prentice Hall.

Fisher's Z description with formulas: <http://davidmlane.com/hyperstat/A50760.html>

Fisher's z' -Transformation, from *MathWorld*—a Wolfram Web resource, by E. W. Weisstein: <http://mathworld.wolfram.com/Fishersz-Transformation.html>

R reference: <http://www.maths.lth.se/help/R/.R/library/SIN/html/fisherz.html>

SAS reference: <http://support.sas.com/ctx/samples/index.jsp?sid=494>

FOURIER TRANSFORM

The Fourier transform takes a function (or sequence) defined in the time or spatial domain and transforms it to the frequency domain, which provides a natural environment for studying many problems. Fourier

analysis (often referred to as spectral analysis) is usually associated with the study of periodic behavior (e.g., sunspots) but is also used to understand nonperiodic and stochastic behavior. Spectral analysis techniques are some of the most ubiquitous tools in modern science and are used in fields as diverse as signal processing, astronomy, geophysics, medical imaging, neurophysiology, speech analysis, and optics.

The Fourier transform can be applied to multidimensional processes; however, it has undoubtedly been applied most widely to one-dimensional processes. Hence, this discussion will refer only to functions (or sequences) defined in the time domain.

Historical Aspects

The Fourier transform is named after the French engineer Jean Baptiste (Joseph Baron) Fourier (1768–1830), who, motivated by his work modeling heat conduction, proposed that any function could be decomposed into a superposition of sinusoidal (sine and cosine) terms. It has since been found that the decomposition is valid only for functions that satisfy certain conditions, which are rather technical (the interested reader is referred to the references). However, almost all functions that arise in physical applications will satisfy the conditions or at least will be well approximated by a sum of sinusoidal terms.

Which Fourier Transform?

Without context, the term *Fourier transform* generally refers to the continuous Fourier transform, which is a linear mapping from a continuous time interval to the frequency domain. The discrete Fourier transform is then the equivalent form for discrete time.

The frequency domain representation allows analysis of a function's frequency characteristics, such as the contribution of the function to

sinusoids at different frequencies. The inverse Fourier transform reverses the transformation from the frequency domain back to the time domain. The mapping is unique, so the inverse Fourier transform will reproduce the original function exactly. The dual transforms are called the *Fourier transform pair*.

What Do We Mean by Frequency?

Frequency is the number of times a function repeats itself within a unit of time. If the time unit is a second, the frequency measure is hertz, the number of cycles per second. The period is the time taken for the function to repeat; in other words, period is the reciprocal of frequency: $\text{Frequency} = 1/\text{Period}$. Frequency can intuitively be considered in terms of sound waves: A bass note is a low-frequency sound, and a whistle is a high-frequency sound.

The period between successive peaks of the solid line in Figure 1 is 8 seconds, giving a frequency of $f = \frac{1}{8}$ Hz. The dotted line is of a higher frequency, $f = \frac{1}{4}$, shown by the period of 4 seconds. The amplitude of a

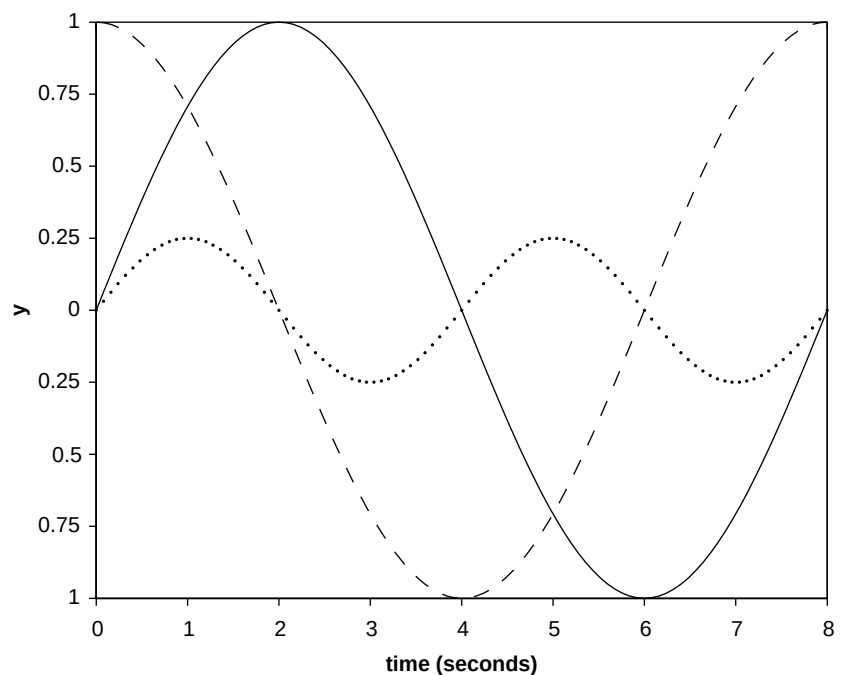


Figure 1 Sinusoidal Functions of Various Frequencies and Phases

Note: Solid line is $\sin(2\pi ft)$ with frequency $f = \frac{1}{8}$, dashed line is $\sin(2\pi t + \pi/2) = \cos(2\pi ft)$, and dotted line is $\sin(2\pi ft)/4$ with twice the frequency, or $f = \frac{1}{4}$.

sinusoid is the height of each peak or trough. The solid line has amplitude one, whereas the dotted line has amplitude $\frac{1}{4}$. The phase of a sinusoid indicates which part of the cycle the function commences at time zero. The solid line has phase 0 (a sine term), whereas the dashed line has the same frequency, but the phase is shifted by $\pi/2$ (equivalent to a cosine term). Hence, cosine functions are simply phase-shifted sine functions.

Continuous Fourier Transform

The continuous Fourier transform $X(f)$ of a function $x(t)$ at frequency f can be expressed as

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt,$$

with corresponding inverse Fourier transform

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{i2\pi ft} df,$$

where i is the square root of -1 . The integrals decompose the function into sinusoidal terms at an infinite division of frequencies. The Fourier transform can be used for complex-valued functions. However, in most applications, the functions are real-valued, which will be assumed in this discussion.

It is not possible to venture into the mathematical detail of the Fourier transform. The interested reader is referred to the references. However, it is useful to consider why the integrand involves complex exponentials rather than sinusoidal terms, as one might expect. Insight comes from Euler's formula:

$$e^{i\vartheta} = \cos(\vartheta) + i\sin(\vartheta).$$

Hence, the complex exponential gives a complex number whose real part is a cosine term and whose imaginary part is a sine term. We know that a sinusoid term with a particular phase angle can be written as a

combination of simple sine and cosine terms. Hence, the complex exponential allows us to consider sinusoidal functions, but in a much more convenient functional form (integrals of exponential functions are easier than sinusoids).

The Fourier transform of a function carries information about the amplitude and phase of contributions at each frequency. The absolute value of the Fourier transform is related to the amplitude. The phase is given by the angle in complex space (called the *argument of the complex number*). A plot of the square of the absolute value of the Fourier transform against frequency is called the *power spectrum of the signal*. The corresponding plot of the phases is the *phase spectrum*. For real functions, the power spectrum is symmetric, and the phase spectrum is antisymmetric, so only the positive frequencies are usually shown.

The Fourier transform has many interesting properties. For example, the Fourier transform of an even function is entirely real and even, whereas for an odd function, it is entirely complex and odd. A time shift changes only the phase, not the amplitude, so the power spectrum is unchanged. The Fourier transform of the sum of two functions is simply the sum of their respective Fourier transforms (known as *linearity*). Examples of a few important functions and their Fourier transforms are shown in Figure 2. A delta function can be thought of as infinity at a single point. The sinc function is defined as $\sin(\pi f)/\pi f$.

One of the most useful properties of the Fourier transform is its behavior under convolution, which makes an efficient computational tool in many scientific problems. Convolution includes operations like smoothing and filtering of a function (signal). The result of the convolution of a function in the time domain is equivalent to multiplication of the Fourier transform of the signal with the transform of the convolution function. Similarly, convolution in the frequency domain is equivalent to multiplication in the time domain. Hence, a convolution (usually a computationally intensive operation) can efficiently be carried out by a simple multiplication in the alternate domain.

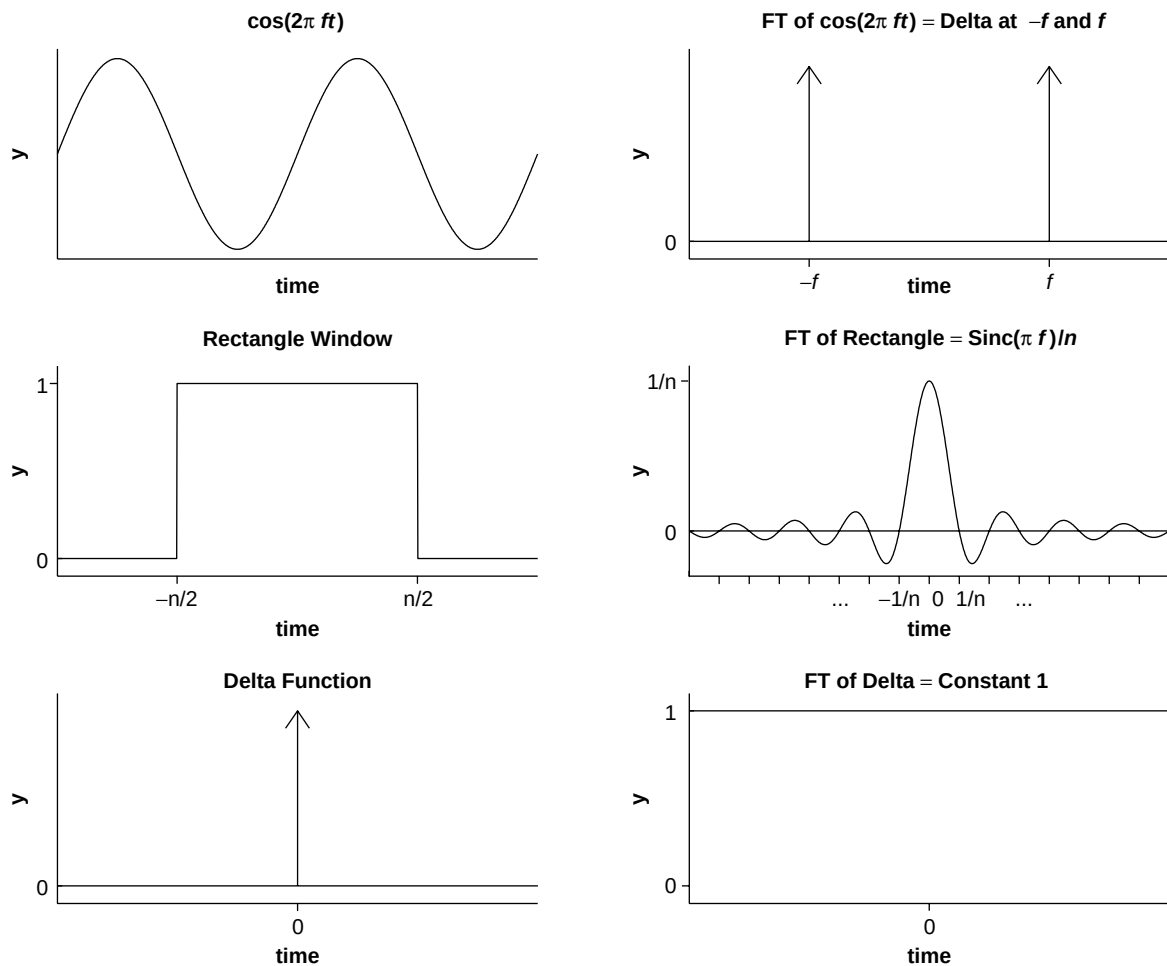


Figure 2 Plots of Various Functions and Their Fourier Transforms

Note: From top to bottom: cosine function, rectangular window, and delta function.

Discrete Fourier Transform

In most practical applications, to enable analysis using computers, a signal is regularly sampled (equally spaced) over a finite time interval. The details associated with irregular sampling are deferred to the interested reader. The sample period, often known as the *fundamental*, is the length of time between the regularly sampled points. The fundamental frequency is then $\frac{1}{\text{Fundamental Period}}$. Integer multiples of the fundamental frequency give the Fourier frequencies. The sampling frequency must be sufficiently small to provide an accurate approximation of

the function, in particular to prevent aliasing, discussed further as follows.

Sinusoids at multiples of the fundamental frequency are called harmonics. A harmonic will complete a whole number of cycles within the time interval; for example, first harmonic completes one cycle, second harmonic two cycles, and so on. Consider the sound produced when a string is struck. The vibrations that produce the sound travel along the string and back again; in other words, the fundamental period is twice the string length. The initial burst of energy produces vibrations at many frequencies. Most frequencies will

die away quickly, their energy being used in producing the sound and some heat. However, the harmonic frequencies will continue to resonate, as they simply bounce back and forth between the ends. The harmonics will slowly die away as energy is lost on producing the sound you hear.

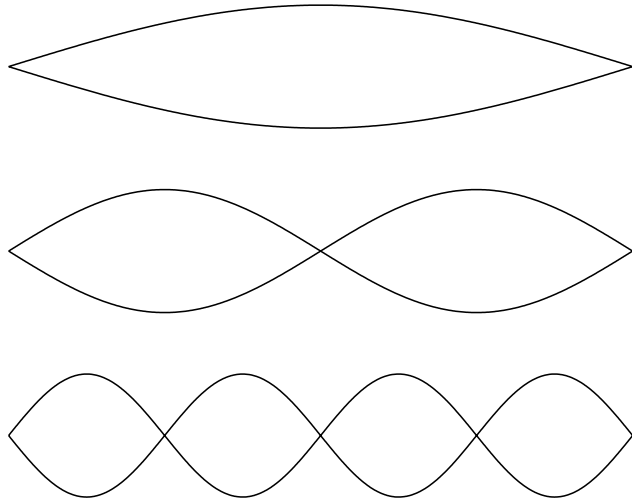


Figure 3 First Three Harmonic Vibrations Along a String, Showing Movement Back and Forth

The discrete Fourier transform takes a sampled signal x_t of finite length n and produces a frequency domain representation X_f ,

$$X_f = \frac{1}{n} \sum_{t=0}^{n-1} x_t e^{-i2\pi ft},$$

with inverse discrete Fourier transform,

$$x_t = \sum_{f=0}^{n-1/n} X_f e^{i2\pi ft}.$$

Usually it is evaluated only at the Fourier frequencies

$$f = -\frac{1}{2}, \dots, -\frac{2}{n}, -\frac{1}{n}, 0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{1}{2}.$$

Some textbooks present the discrete Fourier transform adjusting for different fundamental periods, but this detail is superfluous to this entry.

Essentially, the integral of the continuous Fourier transform is replaced with a summation, which is appropriately normalized. There is no strict convention on whether the normalization ($1/n$) should apply to the discrete Fourier transform or its inverse ($1/\sqrt{n}$) or both. The different normalizations can have a physical meaning in certain contexts. In statistics, it is useful that the power spectrum of the autocovariance function has the same shape as the transform of the original series, but with different phases. This property can enable efficient calculation of the autocorrelation function, often the starting point in time series analysis.

The maximum absolute frequency the discrete Fourier transform can resolve is the Nyquist frequency $1/2$. The signal should be bandlimited so that there is no periodic variation outside of the Nyquist frequency range $(-1/2, 1/2)$. Alternatively, a signal can be subject to an antialiasing filter (called a low-pass filter) prior to sampling, to remove variation outside this range. This is important because of an issue known as aliasing. If a signal has periodicities higher than the Nyquist frequency, then this high-frequency variation is not distinguishable at the sampled points from a signal at lower frequency. Power from all frequencies outside the Nyquist range is folded back into this range. Figure 4 provides an example of sampled sinusoidal signals above the Nyquist frequency indistinguishable from a signal below the Nyquist frequency. Aliasing is the reason stagecoach wheels seem to rotate backwards in some old “western” movies shot at a low frame rate.

As the discrete Fourier transform is usually calculated only at the Fourier frequencies, components between these frequencies are not shown. This problem is known as the *picket fence effect*. For evaluation of the discrete Fourier transform at a higher division of frequencies, the series can be zero-padded (zeros appended to end), which will not impact the power spectrum.

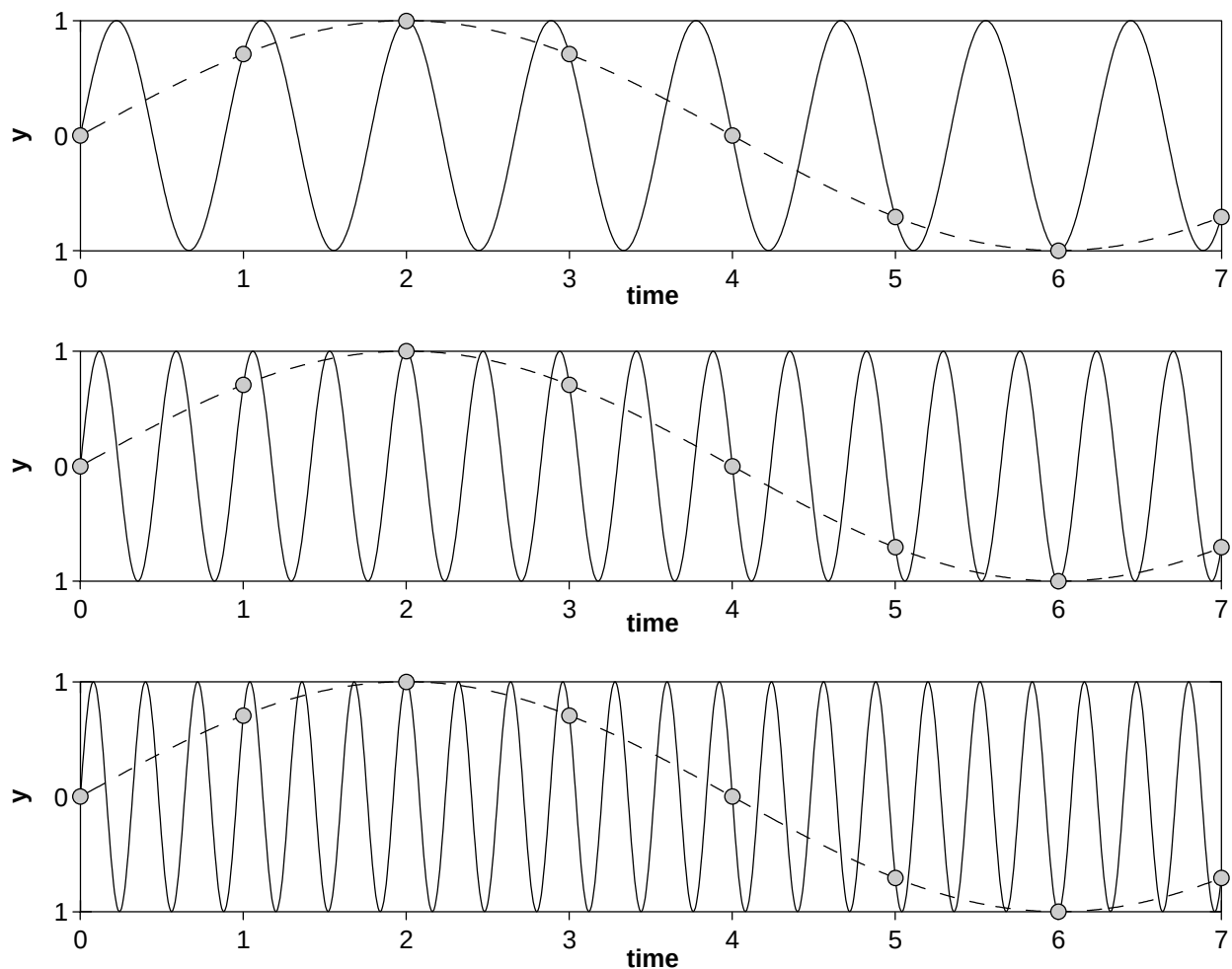


Figure 4 Sinusoidal Functions of Various Frequencies Illustrating Aliasing

Notes: Signal sampled at times $0, \dots, 7$. Dashed line is $\sin(2\pi ft)$ at frequency $f = \frac{1}{8}$. Solid lines (top to bottom) are $\sin[2\pi(f+1)t]$, $\sin[2\pi(f+2)t]$, and $\sin[2\pi(f+3)t]$.

Spectral leakage is caused by observation of a signal of infinite extent over a finite time interval. Observation over a finite interval is equivalent to multiplying the signal by a rectangular window, as in Figure 2. Hence, when the Fourier transform is applied, the resultant series is a convolution of the Fourier transform of the true signal and the transform of the rectangular window, or the sinc kernel shown in Figure 2. The power from each frequency will leak into the rest of the spectrum, a phenomenon known as *spectral leakage*, which can substantially distort the spectrum. It is possible to multiply a series by

an alternative taper window to improve the leakage properties.

The fast Fourier transform is an extremely efficient algorithm for calculating the discrete Fourier transform. The algorithm reduces the number of computations involved from order n^2 to $n \log(n)$. The algorithm requires the length of the series to be a multiple of low-order primes, such as $n = 2^{10} = 1024$.

—Carl J. Scarrott

See also Autocorrelation; Smoothing; Time Series Analysis

Further Reading

- Bloomfield, P. (2000). *Fourier analysis of time series: An introduction*. New York: Wiley.
- Bracewell, R. (1999). *The Fourier transform and its applications* (3rd ed.). New York: McGraw-Hill.
- Percival, D. B., and Walden, A. T. (1993). *Spectral analysis for physical applications: Multitaper and conventional univariate techniques*. Cambridge, UK: Cambridge University Press.
- Shatkay, H. (1995). *The Fourier transform—A primer*. Brown University Technical Report CS-95-37. Retrieved from <http://www.cs.brown.edu/publications/techreports/reports/CS-95-37.html>
- Thibos, L. N. (2003). *Fourier analysis for beginners*. Indiana University School of Optometry Research Library. Retrieved from <http://research.opt.indiana.edu/Library/FourierBook/title.html>
- Fourier transform article: http://en.wikipedia.org/wiki/Fourier_transform
- Fourier transformations information by Wolfram Research: <http://mathworld.wolfram.com/FourierTransform.html>

FRACTAL

There is no universally accepted definition of what constitutes a fractal. However, it is usually clear what one means by a fractal. Fractal objects are objects with strong scaling behavior. That is, there is some relation between the “behavior” of the object at some scale and at finer scales. Figure 1 illustrates some of the *self-similar* geometric fractals (self-similar means that they are similar to pieces of themselves).

IFS Fractals

A particularly nice class of fractal objects is those objects defined by *iterated function systems* (IFSs). An IFS is a formal way of saying that an object is self-similar (i.e., made up of smaller copies of itself). Consider the Sierpinski Gasket, S (shown in Figure 2). We can see that it is made up of three smaller copies of itself. If we think of S as living in $[0,1] \times [0,1]$ and we define the three maps

$$\begin{aligned} w_1(x,y) &= (x/2, y/2), \quad w_2(x,y) \\ &= (x/2 + 1/2, y/2), \quad w_3(x,y) = (x/2, y/2 + 1/2), \end{aligned}$$

then we see that $S = w_1(S) \cup w_2(S) \cup w_3(S)$. The collection of functions $\{w_1, w_2, w_3\}$ make up the IFS. We notice that each w_i is contractive in that for any two points x, y , we see that $d(f(x), f(y)) = (1/2)d(x, y)$, where d measures the usual distance in the plane. Because of this, the combined set mapping $W(B) = w_1(B) \cup w_2(B) \cup w_3(B)$ is also contractive (in the appropriate sense), so by the contraction mapping theorem there is a unique set A with $W(A) = A$. This set A is the *attractor* of the IFS (in this case, the Sierpinski Gasket).

One nice consequence of the contractivity of the IFS mapping W is that we can start with any set B , and the iterates $W^n(B)$ will converge to the attractor of the IFS. As illustrated in Figure 3, after one application of W , we have 3 smaller copies of our original set (a smile in this case), then 9 even smaller copies, then 27 still smaller copies, and on and on until, eventually, all the details of the original set are too small to see and all we see is the overall structure of the Sierpinski Gasket. A moment's thought will convince you that the same thing would happen with any other initial set.

Now we give a more formal definition of a (geometric) IFS. We start with a complete metric space (X, d) and a finite collection of self-maps $w_i: X \rightarrow X$, with $d(w_i(x), w_i(y)) \leq s_i d(x, y)$ for all x, y , where $0 \leq s_i < 1$ is the contraction factor for w_i . Let $H(X)$ denote the set of all nonempty compact subsets of X , and define the metric h (the Hausdorff metric) on $H(X)$ by

$$h(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\}.$$

It turns out that this is a metric on $H(X)$, which makes it into a complete metric space. Furthermore, under this metric, the set map $W: H(X) \rightarrow H(X)$ defined by

$$W(B) = \cup_i w_i(B)$$

is contractive with contractivity factor s .

Geometric fractals are interesting and useful in many applications as models of physical objects, but many times one needs a functional model. It is easy to extend the IFS framework to construct fractal functions.

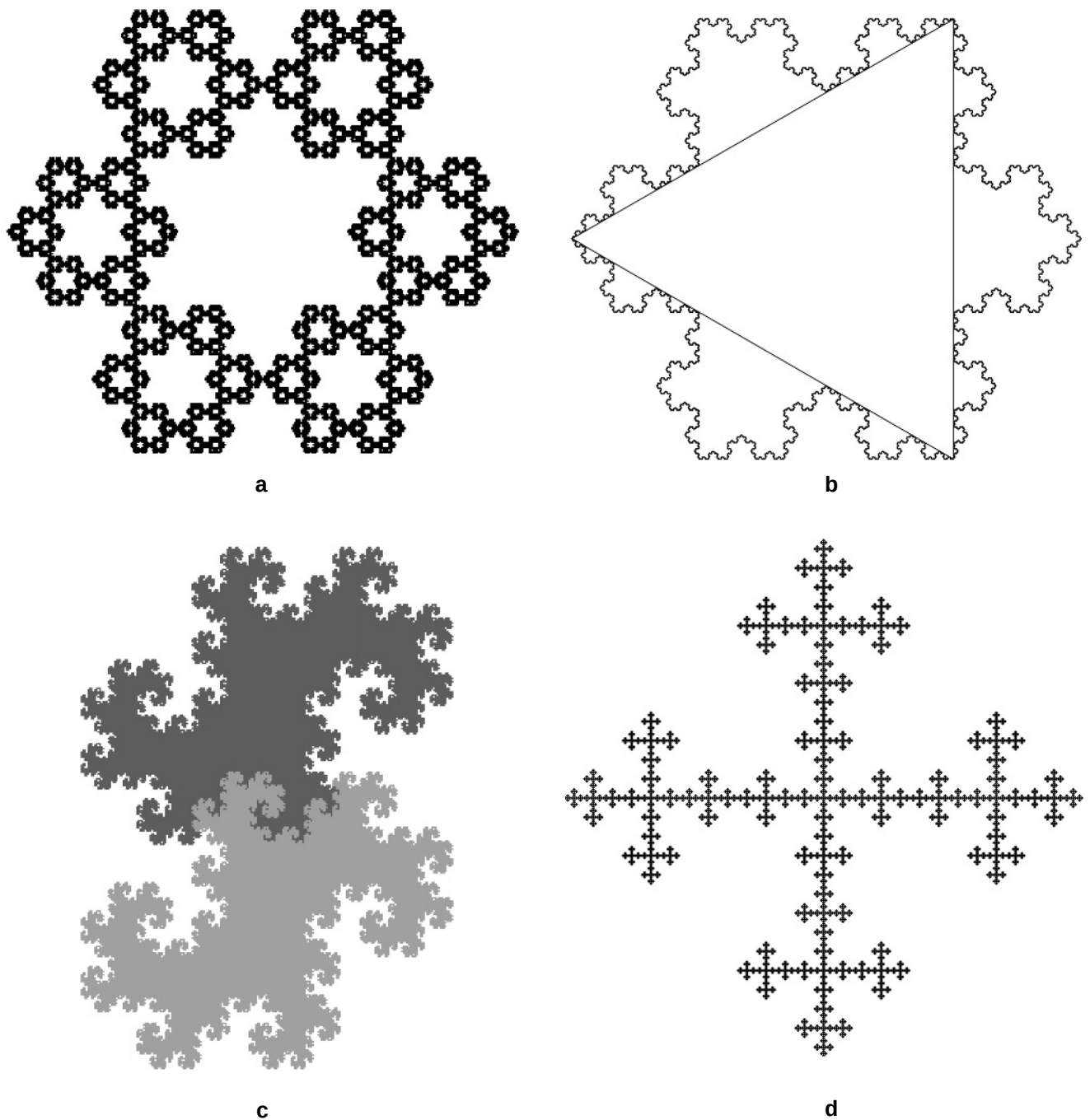


Figure 1 Three Self-Similar Fractal Shapes

Fractal Functions

There are several different IFS frameworks for constructing fractal functions, but all of them have a common core, so we concentrate on this core. We illustrate the ideas by constructing fractal functions on the unit

interval; that is, we construct functions $f : [0,1] \rightarrow \mathbb{R}$. Take the three mappings $w_1(x) = x/4$, $w_2(x) = x/2 + 1/4$, and $w_3(x) = x/4 + 3/4$ and notice that $[0,1] = w_1[0,1] \cup w_2[0,1] \cup w_3[0,1]$, so that the images of $[0,1]$ under each w_i tile $[0,1]$.

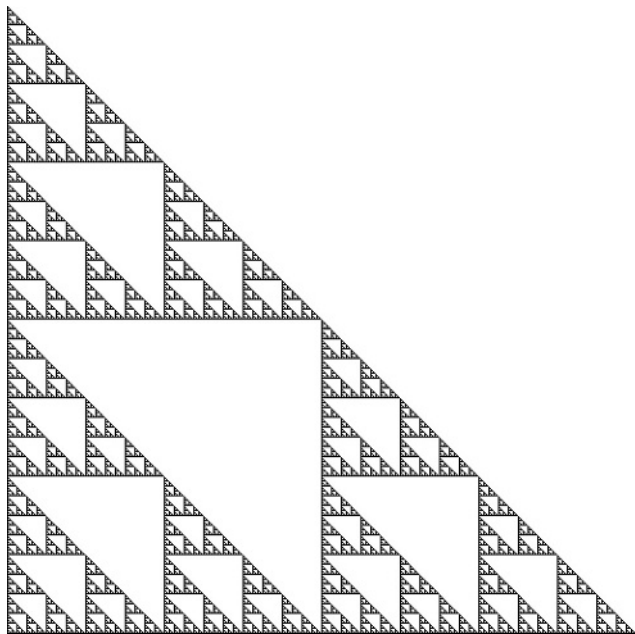


Figure 2 The Sierpinski Gasket

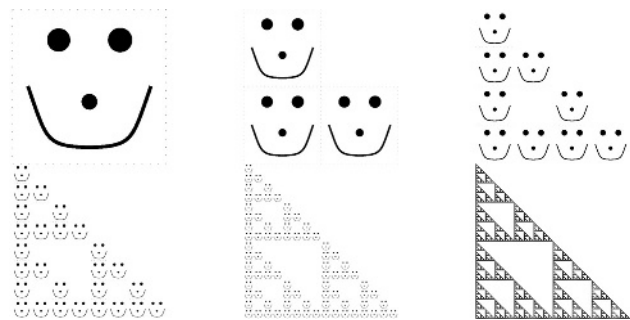


Figure 3 Smiles Converging to the Sierpinski Gasket

Choose three numbers $\alpha_1, \alpha_2, \alpha_3$ and three other numbers $\beta_1, \beta_2, \beta_3$ and define the operator T by

$$T(f)(x) = \alpha_i f(w_i^{-1}(x)) + \beta_i \quad \text{if } x \in w_i([0,1]),$$

where $f: [0,1] \rightarrow \mathbb{R}$ is a function. Then clearly $T(f): [0,1] \rightarrow \mathbb{R}$ is also a function, so T takes functions to functions.

There are various conditions under which T is a contraction. For instance, if $|\alpha_i| < 1$ for each i , then T is contractive in the supremum norm given by

$$\|f\|_{\sup} = \sup_{x \in [0,1]} |f(x)|,$$

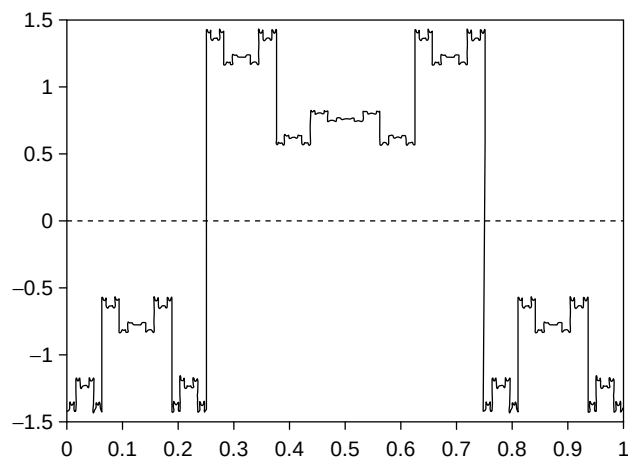


Figure 4 The Attractor of an IFS With Maps

so $T^n(f)$ converges uniformly to a unique fixed point $\bar{f}$ for any starting function f .

Figure 4 illustrates the limiting fractal functions in the case where $\alpha_1 = -\alpha_2 = \alpha_3 = 0.3$ and $\beta_1 = -\beta_2 = \beta_3 = -1$.

It is also possible to formulate contractivity conditions in other norms, such as the L^p norms. These tend to have weaker conditions, so they apply in more situations. However, the type of convergence is clearly different (the functions need not converge pointwise anywhere, for instance, and may be unbounded).

Statistically Self-Similar Fractals

Many times a fractal object is not self-similar in the IFS sense, but it is self-similar in a statistical sense. That is, either it is created by some random self-scaling process (so the steps from one scale to the next are random) or it exhibits similar statistics from one scale to another. The object in Figure 5 is an example of this type of fractal.

These types of fractals are well modeled by random IFS models, where there is some randomness in the choice of the maps at each stage.

Other Types of Fractals

IFS-type models are useful as approximations for self-similar or almost self-similar objects. However, often these models are too hard to fit to a given situation.

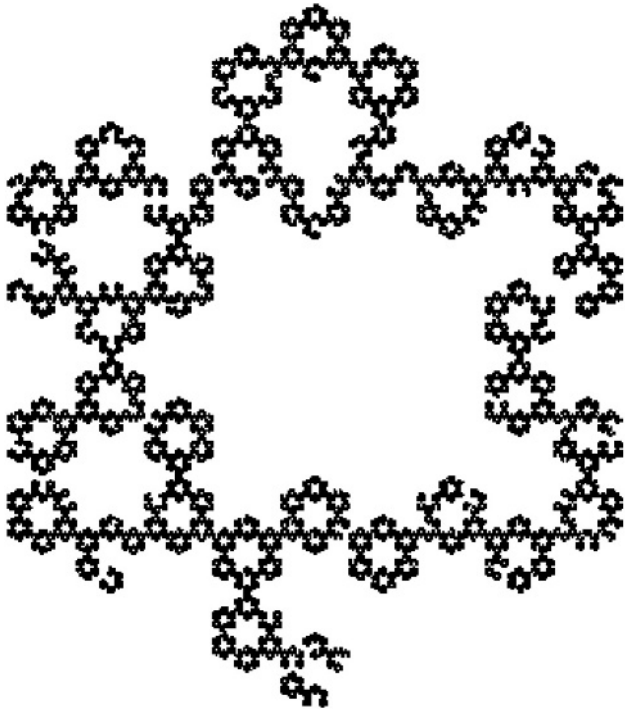


Figure 5 A Statistically Self-Similar Fractal

For these cases, we have a choice—we can either build some other type of model governing the growth of the object, or we can give up on finding a model of the interscale behavior and just measure aspects of this behavior.

An example of the first instance is a simple model for diffusion limited aggregation (DLA), illustrated in Figure 6. In this growth model, we start with a seed and successively allow particles to drift around until they “stick” to the developing object. Clearly the resulting figure is fractal, but our model has no explicit interscale dependence. However, these models allow one to do simulation experiments to fit data observed in the laboratory. They also allow one to measure more global aspects of the model (like the fractal dimension).

Fractal objects also frequently arise as so-called *strange attractors* in chaotic dynamical systems. One particularly famous example is the butterfly-shaped attractor in the Lorentz system of differential equations. These differential equations are a toy model of a weather system and exhibit “chaotic” behavior. The attractor of this system is an incredibly intricate filigreed structure of curves. The fractal nature of this

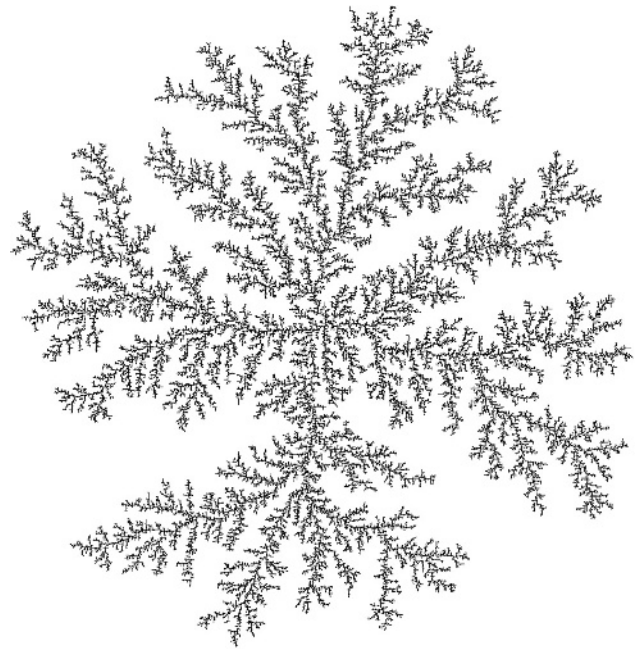


Figure 6 A DLA Fractal

attractor is evident by the fact that, as you zoom in on the attractor, more and more detail appears in an approximately self-similar fashion.

Fractal Random Processes

Since the introduction of fractional Brownian motion (fBm) in 1968 by Benoit Mandelbrot and John van Ness, self-similar stochastic processes have been used to model a variety of physical phenomena (including computer network traffic and turbulence). These processes have power spectral densities that decay like $1/f^a$.

An fBm is a Gaussian process $x(t)$ with zero mean and covariance,

$$E[x(t)x(s)] = (\sigma^2/2)[|t|^{2H} + |s|^{2H} - |t - s|^{2H}],$$

and is completely characterized by the *Hurst exponent* H and the variance $E[x(1)^2] = \sigma^2$. An fBm is statistically self-similar in the sense that for any scaling $a > 0$, we have

$$x(at) = a^H x(t),$$

where by equality we mean equality in distribution. (As an aside and an indication of one meaning of H , the sample paths of an fBm with parameter H are almost surely Hölder continuous with parameter H , so the larger the value of H , the smoother the sample paths of the fBm). Because of this scaling behavior, an fBm exhibits very strong longtime dependence. It is also clearly not a stationary (time invariant) process. This causes many problems with traditional methods of signal synthesis, signal estimation, and parameter estimation. However, wavelet-based methods work rather well, as the scaling and finite time behavior of wavelets match the scaling and nonstationarity of the fBm. With an appropriately chosen (i.e., sufficiently smooth) wavelet, the wavelet coefficients of an fBm become a stationary sequence without the long-range dependence properties. This aids in the estimation of the Hurst parameter H .

Several generalizations of fBms have been defined, including a multifractal Brownian motion. This particular generalization allows the Hurst parameter to be a changing function of time, $H(t)$, in a continuous way. Since the sample paths of an fBm have Hölder continuity H (almost surely), this is particularly interesting for modeling situations in which one expects the smoothness of the sample paths to vary over time.

Fractal Dimension

There are several parameters associated with a fractal object. Of these, fractal dimension (of which there are several variants) is one of the most widely used. Roughly speaking, this “dimension” measures the scaling behavior of the object by comparing it to a power law.

An example will make it more clear. Clearly the line segment $L = [0,1]$ has dimension equal to one. One way to think about this is that if we scale L by a factor of s , the “size” of L changes by a factor of s^1 . That is, if we reduce it by a factor of $1/2$, the new copy of L has $1/2$ the length of the original L .

Similarly the square $S = [0,1] \times [0,1]$ has dimension equal to two since if we scale it by a factor of s , the “size” (in this case, area) scales by a factor of s^2 . How do we know that the “size” scales by a factor

of s^2 ? If $s = 1/3$, say, then we see that we can tile S by exactly $9 = 3^2$ reduced copies of S , which means that each copy has “size” $(1/3)^2$ times the size of the original.

Now take the Sierpinski Gasket S . We see that S is covered by three smaller copies of itself, each copy having been reduced in size by a factor of two. Thus, for $s = 1/2$, we need 3 reduced copies to tile S . This gives

$$\begin{aligned} \text{size of original} &= 3 \times \text{size of smaller copy} \\ \text{copy} &= 3' \times (1/2)^D \text{ size of original copy,} \end{aligned}$$

so $3(1/2)^D = 1$ or $D = \log(3)/\log(2)$. That is, for the Sierpinski Gasket, if we shrink it by a factor of $1/2$, then the “size” gets reduced by a factor of $(1/2)^{\log(3)/\log(2)} = 1/3$. In this sense, the Sierpinski Gasket has dimension $\log(3)/\log(2) \approx 1.5849$, so it has a dimension that is fractional, and so it is a fractal.

We need a different method or definition for the fractal dimension of objects that are not exactly self-similar. The DLA fractal in Figure 6 is a fractal but is not strictly self-similar.

One very common way to estimate the dimension of such an object is the *box counting method*. To do this, cover the image with a square grid (of side length ϵ) and count how many of these boxes are occupied by a point in the image. Let $N(\epsilon)$ be this number. Now repeat this for a sequence of finer and finer grids (letting ϵ tend to 0). We want to fit the relation $N(\epsilon) = a\epsilon^{-D}$, so we take logarithms of both sides to get $\log(N) = \log(a - D\log(\epsilon))$. To estimate D , we plot $\log(N(\epsilon))$ versus $\log(\epsilon)$ and find the slope of the least squares line.

For the object in Figure 6, we have the data in Table 1, which give a fractal dimension of 1.60.

Fractals and Wavelets

We briefly mentioned the connection between fractals and wavelets. Wavelet analysis has become a very useful part of any data analyst’s tool kit. In many ways, wavelet analysis is a supplement (and, sometimes, replacement) for Fourier analysis; the wavelet functions replace the usual sine and cosine basis functions.

Table 1 A Fractal Dimension of 1.60

ϵ	$N(\epsilon)$
1/2	4
1/4	16
1/8	52
2^{-4}	174
2^{-5}	580
2^{-6}	1,893
2^{-7}	6,037
2^{-8}	17,556
2^{-9}	44,399
2^{-10}	95,432

The connection between wavelets and fractals comes because wavelet functions are nearly self-similar functions. The so-called scaling function is a fractal function, and the “mother wavelet” is simply a linear combination of copies of this scaling function. This scaling behavior of wavelets makes it particularly nice for examining fractal data, especially if the scaling in the data matches the scaling in the wavelet functions. The coefficients that come from a wavelet analysis are naturally organized in a hierarchy of information from different scales, and hence doing a wavelet analysis can help one find scaling relations in data, if such relations exist.

Of course, wavelet analysis is much more than just an analysis to find scaling relations. There are many different wavelet bases. This freedom in choice of basis gives greater flexibility than Fourier analysis.

—Franklin Mendivil

Further Reading

- Barnsley, M. G. (1988). *Fractals everywhere*. New York: Academic Press.
- Dekking, M., Lévy Véhel, J., Lutton, E., & Tricot, C. (Eds.). (1999). *Fractals: Theory and applications in engineering*. London: Springer.
- Feder, J. (1988). *Fractals*. New York: Plenum.
- Hutchinson, J. E. (1981). Fractals and self-similarity. *Indiana University Mathematics Journal*, 30, 713–747.
- Lévy Véhel, J., & Lutton, E. (Eds.). (2005). *Fractals in engineering: New trends in theory and applications*. London: Springer.
- Lévy Véhel, J., Lutton, E., & Tricot, C. (Eds.). (1997). *Fractals in engineering*. London: Springer.

Mallat, S. (1999). *A wavelet tour of signal processing*. San Diego, CA: Academic Press.

Mandelbrot, B. (1983). *The fractal geometry of nature*. San Francisco: W. H. Freeman.

Mandelbrot, B., & van Ness, J. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10, 422–437.

Ruelle, D. (1988). *Chaotic evolution and strange attractors: The statistical analysis of time series for deterministic nonlinear systems*. Cambridge, UK: Cambridge University Press.

FRACTIONAL RANDOMIZED BLOCK DESIGN

A fractional randomized block design (also called a randomized block fractional factorial design) reduces the number of treatment combinations that must be included in a multitreatment experiment to some fraction ($\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{9}$, and so on) of the total number of treatment combinations. Consider an experiment with five treatments, denoted by the letters *A*, *B*, *C*, *D*, and *E*. If each treatment has two levels, the number of treatment combinations in the experiment is $2 \times 2 \times 2 \times 2 \times 2 = 32$. By using a $\frac{1}{2}$ or a $\frac{1}{4}$ fractional randomized block design, the number of treatment combinations can be reduced to 16 or 8, respectively. However, the reduction in the size of the experiment comes at a price: Considerable ambiguity may exist in interpreting the results of the experiment. Ambiguity occurs because in the case of a $\frac{1}{2}$ fractional design, two names, called *aliases*, can be given to each source of variation. For example, a sum of squares could be attributed to the effects of treatment *A* and the *BCDE* interaction. In a one-fourth fractional randomized block design, each source of variation has four aliases. Treatments are customarily aliased with higher-order interactions that are assumed to equal zero. This helps minimize but does not eliminate ambiguity in interpreting the outcome of an experiment. One can never be sure that the higher-order interaction is really equal to zero. Because the interpretation of fractional randomized block designs always involves some ambiguity, the designs are most useful for pilot experiments and for exploratory research situations that permit follow-up experiments to be performed. Thus, a large

number of treatments, typically seven or more, can be investigated efficiently in an initial experiment, with subsequent smaller experiments designed to clarify the results or follow up on the most promising independent variables.

A $\frac{1}{2}$ fractional randomized block design in which each treatment has two levels is denoted by 2^{k-1} , where 2^k indicates that each of the k treatments has 2 levels. The -1 in 2^{k-1} indicates that the design is a one-half fraction of a complete 2^k factorial design. This follows because the designation for a one-half fraction of 2^k can be written as $\frac{1}{2}2^k = 2^{-1}2^k = 2^{k-1}$. A one-fourth fractional randomized block design is denoted by 2^{k-2} because $\frac{1}{4}2^k = \frac{1}{2}2^{k-1} = 2^{-2}2^k = 2^{k-2}$.

Procedures for Constructing a Fractional Randomized Block Design

A 2^{5-1} design reduces the number of treatment combinations in an experiment from 32 to 16. The highest-order interaction, $ABCDE$, is typically used to determine which treatment combinations are in the experiment. This interaction, which is called the *defining relation*, divides the treatment combinations into two sets, each containing 16 combinations.

Several schemes have been devised to partition the treatment combinations into orthogonal subsets. One scheme that uses modular arithmetic is applicable to designs of the form p^{k-i} , where i indicates that the design is a $\frac{1}{2}$, $\frac{1}{3}$, and so on, fractional replication and p is a prime number. Let a_j , b_k , c_l , d_m , e_o , z , and p correspond to properties of a design as follows:

a_j , b_k , c_l , d_m , and e_o denote levels of treatments A through E , respectively, where the first level of a treatment is denoted by 0 and the second level is denoted by 1;

$z = 0$ or 1 identifies one of the defining relations: $(ABCDE)_z = (ABCDE)_0$ or $(ABCDE)_1$; and

p denotes the number of levels of each treatment.

The expression $a_j + b_k + c_l + d_m + e_o = z(\text{mod } p)$ says to add the treatment levels (0 or 1) represented by a_j , b_k , c_l , d_m , and e_o and reduce the sum modulo p ; that is, express the sum as a remainder equal to z with

respect to the modulus p . For example, to find a treatment combination in the subset of treatment combinations denoted by $(ABCDE)_0$, find a combination whose sum when divided by p leaves the remainder $z = 0$. Each of the following 16 treatment combinations is in this subset:

00000, 00011, 00101, 01001, 10001, 00110, 01010,
10010, 01100, 10100, 11000, 01111, 11011, 11101,
10111, 11110 = $a_j + b_k + c_l + d_m + e_o = 0(\text{mod } 2)$.

For example, the sum of the treatment levels represented by $0 + 0 + 0 + 1 + 1 = 2$. When 2 is divided by $p = 2$, the remainder is 0. Hence, the treatment combination represented by 00011 satisfies the relation $(ABCDE)_0$. The treatment combinations that satisfy the relation $(ABCDE)_1$ are

00001, 00010, 00100, 01000, 00111, 01011, 01101,
01110, 10000, 11100, 11010, 11001, 10110, 10101,
10011, 11111 = $a_j + b_k + c_l + d_m + e_o = 1(\text{mod } 2)$.

A researcher can flip a coin to decide which of the two sets of 16 treatment combinations to use in an experiment.

A *confounding contrast* is used to assign the 16 treatment combinations in the subset, say, $(ABCDE)_0$, to two groups of blocks. Suppose the researcher chooses CD as the confounding contrast. The following eight treatment combinations satisfy the two relations $(ABCDE)_0$ and $(CD)_0$:

00000, 00110, 01001, 01111, 10001, 10111, 11000,
11110 = $a_j + b_k + c_l + d_m + e_o = 0(\text{mod } 2)$ and
 $c_l + d_m = 0(\text{mod } 2)$.

The remaining eight treatment combinations satisfy the relations $(ABCDE)_0$ and $(CD)_1$:

00011, 00101, 01010, 01100, 10010, 10100, 11101,
11011 = $a_j + b_k + c_l + d_m + e_o = 0(\text{mod } 2)$ and
 $c_l + d_m = 1(\text{mod } 2)$.

This design has two blocks of size eight.

It is often difficult to obtain two blocks of eight experimental units such that the units in a block are

Table 1 Layout for Fractional Randomized Block Design With Five Treatments

	Treatment Combinations			
	$a_0 b_0 c_0 d_0 e_0$ Y_{00000}	$a_0 b_1 c_1 d_1 e_1$ Y_{01111}	$a_1 b_0 c_1 d_1 e_1$ Y_{10111}	$a_1 b_1 c_0 d_0 e_0$ Y_{11000}
Block $(CD)_0 (CE)_1$	$a_0 b_0 c_1 d_1 e_0$ Y_{00110}	$a_0 b_1 c_0 d_0 e_1$ Y_{01001}	$a_1 b_0 c_0 d_0 e_1$ Y_{10001}	$a_1 b_1 c_1 d_1 e_0$ Y_{11110}
Block $(CD)_1 (CE)_0$	$a_0 b_0 c_1 d_0 e_1$ Y_{00101}	$a_0 b_1 c_0 d_1 e_0$ Y_{01010}	$a_1 b_0 c_0 d_1 e_0$ Y_{10010}	$a_1 b_1 c_1 d_0 e_1$ Y_{11101}
Block $(CD)_1 (CE)_1$	$a_0 b_0 c_0 d_1 e_1$ Y_{00011}	$a_0 b_1 c_1 d_0 e_0$ Y_{01100}	$a_1 b_0 c_1 d_0 e_0$ Y_{10100}	$a_1 b_1 c_0 d_1 e_1$ Y_{11011}

Note: Defining relation = $(ABCDE)_0$, confounding contrasts = $(CD)_z$ and $(CE)_z$, generalized interaction = $(DE)_z$.

relatively homogeneous. To reduce the block size from eight to four, two confounding interactions can be used. Suppose a researcher chooses CE as the second confounding contrast. The first block contains the treatment combinations that satisfy three relations, $(ABCDE)_0$, $(CD)_0$, and $(CE)_0$:

$$\begin{aligned} 00000, 01111, 10111, 11000 &= a_j + b_k + c_l + d_m + e_o \\ &= 0(\text{mod } 2), c_l + d_m = 0(\text{mod } 2), \text{ and } c_l \\ &+ e_o = 0(\text{mod } 2). \end{aligned}$$

The second through the fourth blocks contain the combinations that satisfy the relations (a) $(ABCDE)_0$, $(CD)_0$, and $(CE)_1$, (b) $(ABCDE)_0$, $(CD)_1$, and $(CE)_0$, and (c) $(ABCDE)_0$, $(CD)_1$, and $(CE)_1$. The design is shown in Table 1.

When two confounding contrasts are used to reduce the block size, a third interaction or treatment, called the *generalized interaction* or *generalized treatment*, also is confounded with the between-block variation. For the design in Table 1, the generalized interaction is obtained by multiplying the confounding contrasts and reducing the sums of the exponents modulo p , that is, expressing the sums as a remainder with respect to the modulus p . For example,

$$\begin{aligned} CD \times CE &= C^2 D^1 E^1 = 0(\text{mod } 2) \\ &= C^0 D^1 E^1 = D^1 E^1 = DE. \end{aligned}$$

Hence, the generalized interaction of the two confounding interactions is DE .

The alias for a source of variation is obtained by multiplying the label for the source of variation by the defining relation and reducing the sums of the exponents modulo p . For example, the alias for the source

Table 2 ANOVA Table for Fractional Randomized Block Design With Five Treatments

Source	Alias	df
Blocks (CD) , (CE) , (DE)	(ABE) , (ABD) , (ABC)	3
A	$(BCDE)$	1
B	$(ACDE)$	1
C	$(ABDE)$	1
D	$(ABCE)$	1
E	$(ABCD)$	1
AB	(CDE)	1
AC	(BDE)	1
AD	(BCE)	1
AE	(BCD)	1
BC	(ADE)	1
BD	(ACE)	1
BE	(ACD)	1
Error = pooled two- and three-treatment interactions		7
Total		15

Note: Defining relation = $(ABCDE)_0$, confounding contrasts = $(CD)_z$ and $(CE)_z$, generalized interaction = $(DE)_z$.

Table 3 Computational Procedures for Fractional Randomized Block Design With Four Treatments

ABCD Summary Table, Entry is Y_{jklm}											
Block $(AB)_0$	$a_0\ b_0\ c_0\ d_0$	$a_0\ b_0\ c_1\ d_1$	$a_1\ b_1\ c_0\ d_0$	$a_1\ b_1\ c_0\ d_0$	55						
	11	14	12	18							
Block $(AB)_1$	$a_0\ b_1\ c_0\ d_1$	$a_0\ b_1\ c_1\ d_0$	$a_1\ b_0\ c_0\ d_1$	$a_1\ b_0\ c_1\ d_0$	45						
	15	22	5	3							
<i>AB Summary Table</i>		<i>AC Summary Table</i>		<i>BC Summary Table</i>							
<i>Entry is $\Sigma_{l=1}^r Y_{jklm}$</i>		<i>Entry is $\Sigma_{k=1}^q Y_{jklm}$</i>		<i>Entry is $\Sigma_{j=1}^p Y_{jklm}$</i>							
	b_0	b_1		c_0	c_1		c_0	c_1			
a_0	25	37	62	a_0	26	36	62	b_0	16	17	33
a_1	8	30	38	a_1	17	21	38	b_1	27	40	67
	33	67	100		43	57	100		43	57	100

Notes: Y_{jklm} = a score for the experimental unit in treatment combination $a_j b_k c_l d_m$; $j = 1, \dots, p$ levels of treatment A (a_j); $k = 1, \dots, q$ levels of treatment B (b_k); $l = 1, \dots, r$ levels of treatment C (c_l); $m = 1, \dots, t$ levels of treatment D (d_m).

Sums and Sums of Squares

$$\sum_{j=1}^p \sum_{k=1}^q \sum_{l=1}^r \sum_{m=1}^t Y_{jklm} = 11 + 14 + \dots + 3 = 100.00$$

$$\sum_{j=1}^p \sum_{k=1}^q \sum_{l=1}^r \sum_{m=1}^t Y_{jklm}^2 = (11)^2 + (14)^2 + \dots + (3)^2 = 1,528.00$$

$$SSBLOCKS = \frac{(55)^2}{4} + \frac{(45)^2}{4} - \frac{(100.00)^2}{(2)(2)(2)} = 12.50$$

$$SSA = \frac{(62)^2}{(2)(2)} + \frac{(38)^2}{(2)(2)} - \frac{(100.00)^2}{(2)(2)(2)} = 72.00$$

$$SSB = \frac{(33)^2}{(2)(2)} + \frac{(67)^2}{(2)(2)} - \frac{(100.00)^2}{(2)(2)(2)} = 144.50$$

$$SSC = \frac{(43)^2}{(2)(2)} + \frac{(57)^2}{(2)(2)} - \frac{(100.00)^2}{(2)(2)(2)} = 24.50$$

$$SSD = \left[(11)^2 + (14)^2 + \dots + (3)^2 \right] - \left[\frac{(25)^2}{2} + \dots + \frac{(30)^2}{2} \right] - \left[\frac{(26)^2}{2} + \dots + \frac{(21)^2}{2} \right] \\ - \left[\frac{(16)^2}{2} + \dots + \frac{(40)^2}{2} \right] + \left[\frac{(62)^2}{(2)(2)} + \frac{(38)^2}{(2)(2)} \right] + \left[\frac{(33)^2}{(2)(2)} + \frac{(67)^2}{(2)(2)} \right] + \left[\frac{(43)^2}{(2)(2)} + \frac{(57)^2}{(2)(2)} \right] \\ - \frac{(100.00)^2}{(2)(2)(2)} = 2.00$$

$$SSAC = \left[\frac{(26)^2}{2} + \dots + \frac{(21)^2}{2} \right] - \left[\frac{(62)^2}{(2)(2)} + \frac{(38)^2}{(2)(2)} \right] - \left[\frac{(43)^2}{2} + \dots + \frac{(57)^2}{2} \right] + \frac{(100.00)^2}{(2)(2)(2)} = 4.50$$

$$SSBC = \left[\frac{(16)^2}{2} + \dots + \frac{(40)^2}{2} \right] - \left[\frac{(33)^2}{(2)(2)} + \frac{(67)^2}{(2)(2)} \right] - \left[\frac{(43)^2}{(2)(2)} + \frac{(57)^2}{(2)(2)} \right] + \frac{(100.00)^2}{(2)(2)(2)} = 18.00$$

$$SSERROR = SSAC + SSBC = 4.50 + 18.00 = 22.50$$

$$SSTOTAL = 1,528.00 - 1,250.00 = 278.00$$

Note: Defining relation = $(ABCD)_0$, confounding contrast = $(AB)_r$.

of variation labeled treatment A , where the defining relation is $ABCDE$, is

$$\begin{aligned} ABCDE \times A &= A^2B^1C^1D^1E^1 \\ &= 0(\text{mod } 2) = A^0B^1C^1D^1E^1 \\ &= BCDE. \end{aligned}$$

Hence, treatment A and the $BCDE$ interaction are alternative names for the same source of variation. The source of variation represented by the blocks has seven aliases: blocks plus the two confounding contrasts and generalized interaction, plus their aliases. The sources of variation and aliases for the design are shown in Table 2.

Computational Example

The computational procedures for a 2^{4-1} design with $(ABCD)_0$ as the defining relation and AB as the confounding contrast will be illustrated. A fractional factorial design with only four treatments is unrealistically small, but the small size simplifies the presentation. The layout and computational procedures for the design are shown in Table 3.

The analysis is summarized in Table 4. An examination of the table reveals that the $\frac{1}{2}$ fractional design contains the treatment combinations of a complete factorial design with treatments A , B , and C . Treatment D and all interactions involving treatment D are aliased with the sources of variation for the three-treatment design.

Advantages of Fractional Randomized Block Designs

Each source of variation for the design in Table 4 has two labels. You may wonder why anyone would use such a design—after all, experiments are supposed to resolve ambiguity, not create it. Fractional factorial designs are usually used in exploratory research situations where a large number of treatments must be investigated. In such designs, it is customary to limit all treatments to either two or three levels, thereby increasing the likelihood that higher-order interactions are small relative to treatments and lower-order

Table 4 ANOVA Table for Fractional Randomized Block Design With Four Treatments

Source (Alias)	SS	df	MS	F
Blocks or (AB), (CD)	12.50	$n - 1 = 1$	12.50	1.11
A (BCD)	72.00	$p - 1 = 1$	72.00	6.40
B (ACD)	144.50	$q - 1 = 1$	144.50	12.84
C (ABD)	24.50	$r - 1 = 1$	24.50	2.18
D (ABC)	2.00	$t - 1 = 1$	2.00	0.18
ERROR	22.50		11.25	
AC (BD)		$(p - 1)(r - 1)$		
+ BC (AD)		$+ (q - 1)(r - 1) = 2$		
Total	278.00	$pqr - 1 = 7$		

Note: Defining relation = $(ABCD)_0$; confounding contrast = $(AB)_z$.

interactions. Under these conditions, if a source of variation labeled treatment A and its alias, say the $BCDEFG$ interaction, is significant, it is likely that the significance is due to the treatment rather than the interaction. A fractional factorial design can dramatically decrease the number of treatment combinations that must be run in an experiment. An experiment with seven treatments, each having two levels, contains 128 treatment combinations. By the use of a one-fourth fractional factorial design, 2^{7-2} design, the number of treatment combinations in the experiment can be reduced from 128 to 32. If none of the seven-treatment F statistics are significant, the researcher has answered the research questions with one fourth the effort required for a complete factorial design. On the other hand, if several of the F statistics are significant, the researcher can follow up with several small experiments to determine which aliases are responsible for the significant F statistics. Many researchers would consider ambiguity in interpreting the outcome of the initial experiment a small price to pay for the reduction in experimental effort.

—Roger E. Kirk

See also Factorial Design; Multivariate Analysis of Variance (MANOVA)

Further Reading

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

FREQUENCY DISTRIBUTION

Frequency distribution graphs present all the actual data for a single variable. Their purpose is to illustrate the shape and distribution of the data, making it easier to identify outliers, gaps, clusters, and the most common data points.

Stem-and-leaf plots are examples of frequency distributions. Each number in the data set is divided into a stem and a leaf. The stem consists of the first digit or digits, and the leaf consists of the last digit. The stem can have any number of digits, but the leaf will contain only one number. For example, the number 1004 would be broken down into 100 (stem) and 4 (leaf), and the number 1.9 would be broken down into 1 (stem) and 9 (leaf).

Table 1 is a stem-and-leaf plot of grades in a history course. The graph was created using Word (a vertical bar was typed on each line). The plot illustrates that there are 21 grades, one grade of 39, none in the 40s, one grade of 57, two of 63, one of 64, one of 65, one of 67, three of 70, and so on.

Table 1 Final Grades for Students in a History Course

3	9
4	
5	7
6	33457
7	0001122356
8	145
9	4

To create a stem-and-leaf plot, follow these steps:

1. Put all the raw data in numerical order.
2. Separate each number into a stem and a leaf.

—Adelheid A. M. Nicol

See also Cumulative Frequency Distribution; Histogram; Stem-and-Leaf Display

Further Reading

Friel, S. N, Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32, 124–158.

FRIEDMAN TEST

The Friedman test is a rank-based, nonparametric test for several related samples. This test is named in honor of its developer, the Nobel laureate and American economist Milton Friedman, who first proposed the test in 1937 in the *Journal of the American Statistical Association*. A researcher may sometimes feel confused when reading about the Friedman test because the “related samples” may arise from a variety of research settings. A very common way to think of Friedman’s test is that it is a test for treatment differences for a randomized complete block (RCB) design. The RCB design uses blocks of participants who are matched closely on some relevant characteristic. Once the blocks are formed, participants within each block are assigned randomly to the treatment conditions. In the behavioral and health sciences, a common procedure is to treat a participant as a “block,” wherein the participant serves in all the treatment conditions of an independent variable—also commonly referred to as a repeated measures design or a within-subjects design.

Although it is seen relatively rarely in the research literature, there is another research situation in which the Friedman test can be applied. One can use it in the context in which one has measured two or more comparable (also referred to as “commensurable”) dependent variables from the same sample, usually at the same time. In this context, the data are treated much like a repeated measures design wherein the commensurable measures are levels of the repeated measures factor.

There is an additional source of confusion when one thinks about the Friedman test for repeated measures designs because for the RCB design, the parametric method for testing the hypothesis of no differences between treatments is the two-way ANOVA, with treatment and block factors. The Friedman test, which depends on the ranks of the dependent variable within each block, may therefore be considered a two-way ANOVA on ranks.

It is known in theoretical statistics that the Friedman test is a generalization of the sign test and has similar modest statistical power for most distributions that are likely to be encountered in

behavioral and health research. For normal distributions, the asymptotic relative efficiency (ARE) of the Friedman test with respect to the F test, its counterpart among parametric statistical tests, is $0.955k/(k+1)$, where k is the number of treatment groups. When $k = 4$, the ARE of the Friedman test relative to the F test is 0.764. There is evidence from computer simulation studies that the ARE results, which are for very large sample sizes, are close to the relative efficiency to be expected for small and moderate sample sizes.

It is useful to note that in our desire to select the statistical test with the greatest statistical power, we often select the test with the greatest ARE. Therefore, for a normal distribution, the parametric test is more statistically powerful, leading us to recommend the ANOVA F test over the Friedman test. However, in the case of nonnormal distributions of dependent variables, the recommendation favors the Friedman test. Given the frequency at which nonnormal distributions are encountered in research, it is remarkable that nonparametric tests, such as the Friedman test, or other more-powerful tests, such as (a) the Zimmerman-Zumbo repeated measures ANOVA on ranks, wherein the scores in all treatment groups are combined in a single group and ranked, or (b) the Quade test, which uses information about the range of scores in the blocks relative to each other, are not used more often.

The Case Study and the Data

In the field of art education, there is a great deal of interest in whether parents participate in art activities with their young children and whether this participation changes over the early school years. It is often noted in the educational literature that talking with children about book illustrations, providing writing materials at home, and having children try various forms of expression such as drawing and painting enable children to express their creativity and can lead to their using artwork as material for instruction, play, and creative display.

For the purposes of demonstrating the Friedman test, 20 children (11 boys, 9 girls), average age 67.6 months, were selected from the Early Childhood Longitudinal Study, Kindergarten (ECLS-K) class of 1998–99. The ECLS-K focuses on children's early school experiences, collecting information from children and their parents, teachers, and schools. The question "How often do you help your child do art?"—with four response options: (a) *not at all*, (b) *once or twice per week*, (c) *3 to 6 times per week*, and (d) *every day*—was asked of parents when their child was in kindergarten, Grade 1, and Grade 3. The statistical software package SPSS version 13 was used for the analyses. The data are listed in Table 1 by each child's gender, age in months at the kindergarten assessment, and three parent responses to the question "How often do you help your child do art?"

The Assumptions Underlying the Friedman Test

The two assumptions for the test are stated here in terms of the RBC design:

Table 1 Raw Data From the ECLS-K Study

<i>Gender</i>	<i>Age in Months</i>	<i>Kindergarten</i>	<i>Grade 1</i>	<i>Grade 3</i>
Male	67.40	3	2	3
Male	67.40	4	2	3
Female	64.33	2	2	2
Female	64.40	2	2	2
Female	75.20	3	2	2
Male	67.20	4	2	2
Male	70.47	3	3	3
Female	69.03	1	2	3
Male	67.30	3	3	2
Female	75.17	3	2	2
Male	68.47	2	2	1
Male	62.70	2	2	2
Male	65.80	4	3	3
Female	74.47	2	3	2
Female	66.10	3	2	2
Female	68.23	3	3	2
Female	69.17	3	4	4
Male	61.83	2	1	1
Male	62.77	2	2	1
Male	65.33	3	2	2

- The dependent variable is quantitative in nature (ordinal, interval, or ratio scale) so that the rank transformation can be applied.
- The scores within one block do not influence the scores within the other blocks. This is an assumption of statistical independence across blocks. Violation of this assumption will seriously inflate the Type I error rates of the hypothesis test.

The Research Hypothesis

For our example, the “treatment” condition, in RCB design notation, is the grade (i.e., kindergarten, Grade 1, and Grade 3). The most general statistical hypothesis is:

H_0 : The treatment conditions have identical effects. That is, each ranking of the dependent variable within a block is equally likely.

H_a : At least one of the treatment conditions tends to yield larger or smaller observed values than at least one other treatment.

The Friedman test is both an unbiased and a consistent test when testing these hypotheses. If one is willing to make additional assumptions about the distributions of dependent variables, for example that they are symmetric, the Friedman test can also test the equality of mean ranks of the treatments.

To compute the Friedman test, the scores within each block are compared with each other, and the rank of 1 is assigned to the smallest value, the rank of 2 to the next smallest, and so on. Average ranks are used in the case of ties. The computations for the Friedman test are then based on the sum of ranks for each treatment. The resulting test statistic does not follow a regularly shaped (and known) probability distribution, and so an approximation is usually used. The approximate distribution for the Friedman statistic is the chi-square distribution with $k - 1$ degrees of freedom. For this example, SPSS was used to compute the Friedman test of equal mean ranks. The SPSS output below shows that the mean ranks in kindergarten, Grade 1, and Grade 3 are not equal, $\chi^2(2) = 8.32$, $p < .05$.

Although it is beyond the scope of this entry, a post hoc multiple comparison procedure is available. Also, although we demonstrated the Friedman test using the

Ranks	
	Mean Rank
how often parent helps child with art - kindergarten	2.40
how often parent helps child with art - first grade	1.90
how often parent helps child with art - third grade	1.70

Test Statistics ^a	
N	20
Chi - Square	8.320
df	2
Asymp. Sig.	.016

a. Friedman Test

Figure 1 SPSS Results Based on the ECLS-K Study Data

SPSS software, it is also available in other statistical packages, such as SAS, Minitab, and S-Plus.

—Bruno D. Zumbo

See also Inferential Statistics

Further Reading

- Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis*, 42, 569–593.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675–701.
- Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.
- Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *Journal of Experimental Education*, 62, 75–86.

Friedman's Test Applet (allows you to enter data and calculate the test statistic): <http://www.fon.hum.uva.nl/Service/Statistics/Friedman.html>

G

A mathematician is a device for turning coffee into theorems.

—Paul Erdős

GALTON, SIR FRANCIS (1822–1911)

A prodigy, Sir Francis Galton was born of two important families. His mother was Charles Darwin's aunt and the daughter of Erasmus Darwin. His father's line included wealthy bankers and gunsmiths. By age six, Galton was conversant with the *Iliad* and the *Odyssey*. At age seven, he read Marmion's, Cowper's, Pope's, and Shakespeare's works for pleasure. After reading a page twice, he could repeat the text verbatim.

After studying medicine and mathematics at Cambridge and failing to excel in either discipline, Galton took a poll degree in 1843. After his father died, Galton pursued his passion for exploration. In so doing, he became a published author, received gold medals from two geographic societies, and was elected to the Royal Society. Galton developed a method for mapping atmospheric circulation and was the first to recognize the effects of high-pressure systems on weather. Another of his achievements was research proving that one's fingerprints are unique.

In statistics, Galton showed that the law of error could be used not only to estimate true scores, but also to investigate populations in terms of individuals' deviations from the mean. He developed the concept

of regression, which allowed for the first scientific attempts to study the relationship of heredity and human behavior. Galton measured the heights of children and their parents. He plotted the data in such a fashion that the initial regression line decreased the error in prediction. Using sweet pea samples selected on the basis of the weight of parent seeds, he discovered that the progeny seed weights reverted (i.e., regressed) toward the mean of the parent distribution. This discovery led to the development of the concept of correlation. The statistical concepts of regression toward the mean and correlation allowed the field of psychometrics to move forward significantly. Researchers were able to study the stability of human attributes. For instance, through the use of regression and correlation, the relationship between the intelligence scores of parents and their children could be examined.

In *Hereditary Genius* (1869), Galton contended that intellect is inherited. To collect evidence to test his theory, he founded an anthropometric laboratory through which he measured an array of physical attributes (e.g., height, reaction time, and head circumference). Although his notion was fatally flawed, throughout his life Galton argued that nature dominates nurture in the development of human mental ability.

Galton used Darwin's comments on the selective breeding of plants and animals to suggest that humans could be improved by similar methods. In 1883, he coined the term for such practices, *eugenics*, and a long and bitter controversy ensued. Although his name is often linked to negative eugenics (e.g., Hitler's attempts to exterminate "inferior races"), for the most part Galton favored positive approaches (e.g., "genetically superior" people should marry early and produce more children). Nevertheless, he urged that "undesirables" be restricted from free reproduction, even supporting their sterilization.

—Ronald C. Eaves

See also Correlation Coefficient; Intelligence Quotient; Intelligence Tests

Further Reading

Brookes, M. (2004). *Extreme measures: The dark visions and bright ideas of Francis Galton*. London: Bloomsbury.

Sir Francis Galton: <http://www.mugu.com/galton/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, 62(2), 227–240.

Depending on how one interprets what an effect size index is, it may be claimed that its history started around 1940, or about 100 years prior to that with the ideas of **Sir Francis Galton**, Charles Darwin (Galton's cousin), and others. An attempt is made in this article to trace histories of a variety of effect size indices. Effect size bases discussed pertain to (a) relationship, (b) group differences, and (c) group overlap. Multivariate as well as univariate indices are considered in reviewing the histories.

GAMBLER'S FALLACY

The *gambler's fallacy* is a common invalid inductive inference. It involves the mistaken intuition or belief

that the likelihood of a particular outcome of a process that generates independent random events increases as a function of the length of a run of consecutive non-occurrences of that outcome.

For example, a person playing a casino roulette wheel would commit the gambler's fallacy if he or she had a greater tendency to gamble on red than on black after four consecutive black outcomes, than after a shorter run of black outcomes. Such a tendency, or belief that red is more likely to occur as a function of its nonoccurrence, is erroneous, because the outcomes of the spins of a properly calibrated roulette wheel are independent, and the probabilities of red and black are equal and remain constant from spin to spin of the wheel. Similarly, a flip of a fair coin is not more likely to produce tails after a run of heads; nor is a pregnant woman more likely to give birth to a girl if she has, in the past, given birth to three boys consecutively.

The most widely cited explanation of the gambler's fallacy effect involves the hypothesis that people judge the randomness of an observed series of outcomes in terms of the extent to which it represents the output that would be expected of a prototypical random process—one that contains few orderly sequences such as long runs, symmetries, or strict alternations of one outcome, and few over- or under-representations of possible outcomes. Perhaps the gambler's fallacy arises because the occurrence of a locally less frequent outcome would produce a sample that would better represent randomness than the alternative sample would. For example, given five flips of a fair coin, heads might seem more likely after a series such as *THTTT*, because *THTTTH* has a shorter run of tails, and overrepresents tails less, than *THTTTT* does. People may also believe that a random device is somehow capable of correcting for the local scarcity of one outcome by overproducing instances of that outcome. Such thinking is faulty. A random device has no memory or means by which to correct its output, or to prevent patterns from appearing in a sample of outcomes.

Generalization from frequently encountered cases involving finite populations sampled without replacement could also explain this fallacy. For example,

a motorist who is stopped at a railroad crossing waiting for a freight train to pass would be using sound reasoning if he or she counted freight cars that have passed the crossing and compared this number to his or her knowledge of the finite distribution of train lengths to determine when the crossing will clear. However, such reasoning is invalid when applied to large populations sampled without replacement.

—David M. Boynton

See also Law of Large Numbers

Further Reading

- Gold, E. (1997). *The gambler's fallacy*. Unpublished doctoral dissertation, Carnegie Mellon University.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Johnson, J., & Tellis, G. J. (2005). Blowing bubbles: Heuristics and biases in the run-up of stock prices. *Journal of the Academy of Marketing Science*, 33(4), 486–503.

Ads of stocks and mutual funds typically tout their past performance, despite a disclosure that past performance does not guarantee future returns. Are consumers motivated to buy or sell based on past performance of assets? More generally, do consumers (wrongly) use sequential information about past performance of assets to make suboptimal decisions? Use of this heuristic leads to two well-known biases: the hot hand and the **gambler's fallacy**. This study proposes a theory of hype that integrates these two biases—that a positive run could inflate prices, and a negative run could depress them, although the pattern could reverse on extended runs. Tests on two experiments and one event study of stock purchases strongly suggest that consumers dump “losers” and buy “winners.” The latter phenomenon could lead to hyped-up prices on the stock market for winning stocks. The authors discuss the managerial, public policy, and research implications of the results.

GAUSS, CARL FRIEDRICH (1777–1855)

Gauss, who was born April 30, 1777, in Brunswick, Germany, the child of a poor family, was a prodigy who became famous for his advances in many branches of science, but he was, above all else, a mathematician.

Mathematics came naturally to Gauss, who is said to have corrected his father's wage calculations when he was aged three. Early evidence of his ability was provided by his speedy answer to the request by his schoolteacher to sum the numbers from 1 to 100. He saw immediately that each of $(1 + 100)$, $(2 + 99)$, . . . summed to 101 and thus the answer was 5050. He had invented the formula for himself, and this was typical of his early career—he was frequently inventing results and discovering later that these results had been found before. Gauss's ability came to the attention of the Duke of Brunswick, who became his patron and sent him, in 1795, to the University of Göttingen. It was only a matter of time before Gauss outdid his predecessors: In 1796, he showed that it would be possible, using ruler and compasses only, to construct a regular figure with

$$2^{2^n} + 1$$

sides for any integer n .

In 1801, Gauss summarized his discoveries of the previous 5 years in *Disquisitiones Arithmeticae*, a masterpiece that immediately established him as the foremost living mathematician. Turning to astronomy, Gauss next developed new methods for calculating the motions of planets. This was spectacularly confirmed by his ability to predict where astronomers should search for the minor planet Ceres. Gauss's achievement was based on his development of the method of least squares and the use of an error distribution now known variously as the normal or Gaussian distribution. The procedures were set out in his 1809 work *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*.

In 1805, Gauss was happily married to Johanna Ostoff, and he soon had a son and daughter. In 1807, the family moved to Göttingen, where Gauss was

appointed director of the observatory. In 1809, Johanna died in childbirth, and the following year, Gauss married her best friend, Minna Waldeck, although it appears that this marriage was not a happy one.

Gauss's astronomical work resulted in inventions such as the heliotrope, an instrument for accurate direction-finding by means of reflected sunlight. Gauss also experimented with magnetometers; photometers; and, some 5 years before Samuel Morse, the telegraph. He died peacefully at Göttingen in the early morning of February 23, 1855.

—Graham Upton

See also Normal Curve

Further Reading

Dunnington, G. W. (2004). *Carl Friedrich Gauss: Titan of science*. Washington, DC: Mathematical Association of America.

Carl Friedrich Gauss article: http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

GENERALIZED ADDITIVE MODEL

Estimating the linear model $Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + e_i$ is at the core of many of the statistics conducted today. If you allow the individual X variables to be products of themselves and other variables, the linear model is appropriate for factorial ANOVAs and polynomial regressions, as well as estimating the mean, t tests, and so on. The flexibility of the linear model has led authors of some textbooks and software to call this the *general linear model*. I try to avoid this phrase because it can be confused with the *generalized linear model*, or GLM. The GLM is an important extension that allows researchers to analyze efficiently models where the responses are proportions and counts, as well as other situations. More on this later.

The main focus of this entry is extending the linear model into an additive model. In the linear model, each X variable is multiplied by a scalar, the β value. This is what makes it a linear model, but this restricts

the relationship between X and Y (conditioned on all the other X s). With additive models, the β values are replaced by usually fairly simple (in terms of degrees of freedom) functions of the X variables. The model can be rewritten as $Y_i = \alpha + f_1(X_{1i}) + \dots + f_k(X_{ki}) + e_i$. The functions are usually assumed to be splines with a small number of knots. More complex functions can be used, but this may cause the model to overfit the observed data and thus not generalize well to new data sets. The typical graphical output shows the functions and the numeric output shows the fit of the linear and nonlinear components. The choice of functions, which often comes down to the type and complexity of the splines, is critical.

To illustrate this procedure, Berndt's 1991 data from 534 respondents on hourly wages and several covariates (experience in years, gender, and education in years) are considered. One outlier with an hourly wage of \$44 ($z = 6.9$) is removed, but the data remain skewed (1.28, $se = 0.11$). Logging these data removes the skew (0.05, $se = 0.11$), so a fairly common approach is to use the logged values as the response variable and assume that the residuals are normally distributed. Suppose the researchers' main interests are in the experience variable, and whether income steadily increases with experience or whether it increases rapidly until some point and then increases but less rapidly. For argument's sake, let us assume that the increases are both linear with the logged wages. The researchers accept that wages increase with education and believe that the relationship is nonlinear, and so they allow this relationship to be modeled with a smoothing spline. Because the variable *female* is binary, only a single parameter is needed to measure the difference in earnings between males and females. Although categorical variables can be included within generalized additive models (GAMs), the purpose of GAMs is to examine the relationships between quantitative variables and the response variable. The first model is

$$\ln \text{wages}_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{Exper}_i + f_1(\text{Educ}_i) + e_i.$$

This is like a normal multiple linear regression for the variables *female* and *Exper*; the model fits both as

conditionally linear with the log of wages, but the relationship for education is allowed to be curved. This was fit with the gam package for S-Plus with the default spline (smoothing spline with $df = 4$). The residual deviance is 103.74. There is a positive linear relationship between experience and the log of wages and a positive curved relationship between education and the log of wages. The effect for *female* is negative, meaning that after controlling for experience and education, females earn less than their male counterparts. The top three plots in Figure 1 show this model. With GAMs, people usually rely on plots to interpret the models and compare the deviance values, or they use methods of cross-validation to decide how complex the model (including the complexities of the individual f s) should be. Here, the deviance values will be compared.

The bottom three plots of Figure 1 show the model

$$\ln \text{wages}_i = \beta_0 + \beta_1 \text{female}_i + f_1(\text{Exper}_i) + f_2(\text{Educ}_i) + e_i.$$

f_1 has been set to a piecewise linear model, so two lines are connected at a knot determined by the algorithm. The residual deviance drops to 97.55, which is statistically significant ($\chi^2(1) = 6.20, p = .01$). The package used (gam) allows different types of curves (including loess) to be included in the model, although the efficiency of the algorithm works best if the same type is used. What is clear from this model is that a single linear term of experience is not sufficient to account for these data. If we allow the relationship between experience and the log of wages to be a $df = 4$ spline, the

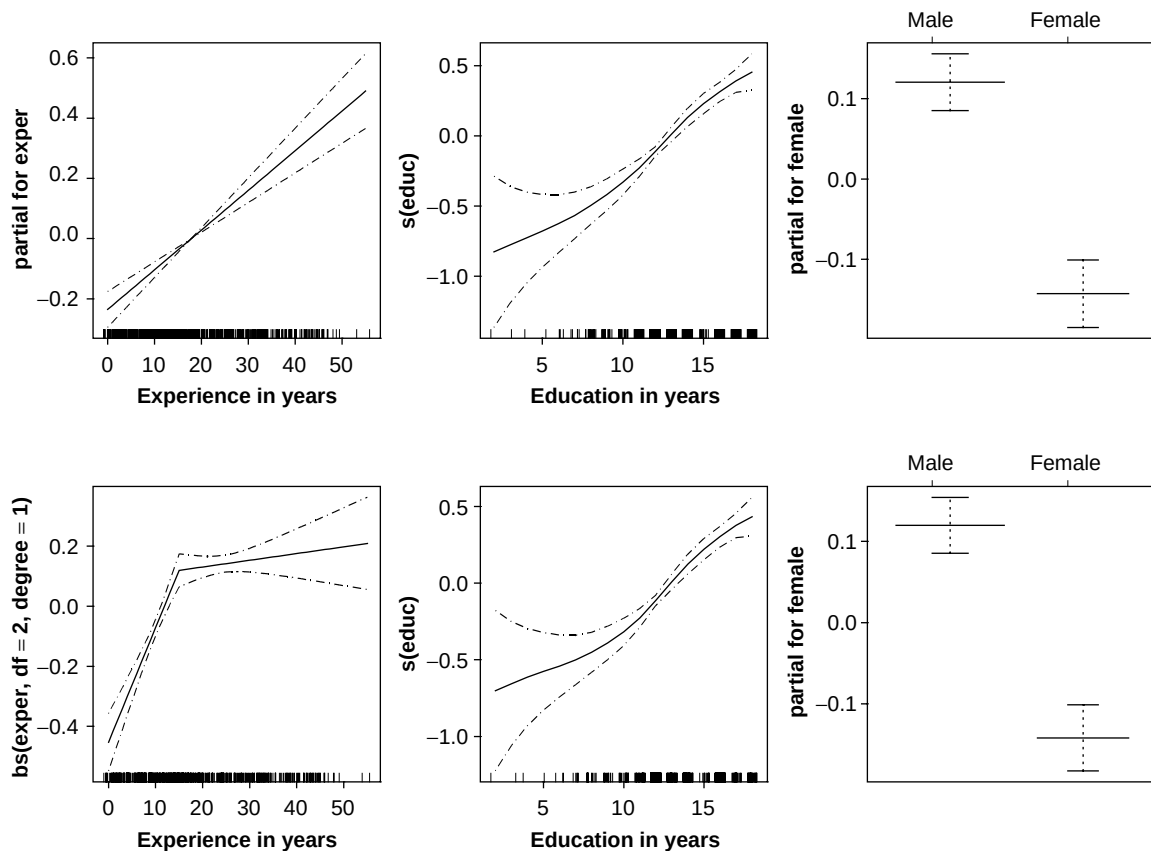


Figure 1 Plots for the Predictor Variables for Two Models Predicting the Log of Wages Using Experience, Education, and Gender

Notes: The top row shows the GAM where experience is a linear predictor. The bottom row has experience as a piecewise linear relationship with a single knot estimated at approximately 15 years' experience. The dashed lines are for standard errors and rugplots are used to show the univariate distributions of experience and education.

residual deviance drops only to 96.85. Although this is not an improvement in the fit of the model, the researcher might still opt for this unless he or she has a theoretical explanation for the sudden change in slope for the previous model.

Just as the generalized linear model allows researchers to model proportions and counts efficiently using the basic concepts of the linear model, the generalized additive model also allows this. The user chooses an error distribution from the exponential family and an associated link function, denoted $g()$. Popular distributions are the normal distribution (associated link is the identity function), the binomial distribution (associated link is the logit function), and the Poisson distribution (associated link is the natural log, or $\ln$, function). In fact, the above example could have been modeled with the log link and assuming Poisson error, and this approach shows that a smooth spline does fit experience better than the piecewise linear model. Generalizing the additive model can be done in the same way as generalizing the linear model. If μ_i is the value of the response variable, then $E(g(\mu_i)) = \eta_i$, where η_i is an additive model of the form $\sum f_k(X_{k_i})$ where one of the X variables is a constant so that an intercept is included. Including the error term, this is $g(\mu_i) = \eta_i + e_i$, where the e_i are assumed to follow some distribution. The wages example could be fit with the following GAM:

$$\begin{aligned}\ln(\mu_i) &= \eta_i + e_i \\ \eta_i &= \alpha + f_1(\text{Exper}_i) + f_2(\text{Educ}_i) + \beta_1 \text{female}_i \\ e_i &\sim \text{Poisson}(\mu_i).\end{aligned}$$

To illustrate a logistic additive model, Vrijji's 2005 data inspired by truth and lie detection using criteria-based content analysis (CBCA) will be used. This is a method used in several countries to try to determine whether a child is telling the truth or a lie when questioned, usually in connection with cases of child sexual abuse. There are 19 criteria, and each statement can be given a 0, 1, or 2. These are summed so that each person can get a score from 0 to 38, with high scores indicating more truthfulness. One problem with this procedure is that people with more linguistic skills tend to have higher scores than people with fewer linguistic

skills. Because of this, there is assumed to be a complex relationship between age, CBCA score, and truth.

Suppose there are 1,000 statements from people who are 3 to 22 years old. All of the statements have CBCA scores, and it is known whether or not they are truthful. For these data, age and truth were created independently, so age, on its own, does not predict truth ($t(998) = 1.04$, $p = .30$). Three GAMs were estimated. The first has just CBCA to predict truth. This uses the logit link function and assumes binomial variation. The default smoothing function for the gam package is used, and the result (in the upper left hand corner of Figure 2) shows that the probability of truth increases with CBCA scores. The deviation from linear is statistically insignificant ($\chi^2(3) = 44.18$, $p < .01$). The residual deviance of this model is 1251.65.

The next model has $\eta_i = \alpha_i + f_1(\text{CBCA}_i) + f_2(\text{age}_i)$, where both f_1 and f_2 are $df = 4$ smoothing splines, and the resulting curves are shown in the second row of Figure 2. CBCA is again positively related to truth. However, age is negatively related (because it is conditional on CBCA). Both curves show marked nonlinearity. The residual deviance is 1209.60, which is a large improvement in fit on the previous model ($\chi^2(3.82) = 42.05$, $p < .001$). The final row in Figure 2 shows the GAM, which includes an interaction term. The graph of the interaction effect (new residual deviance 1201.51, change $\chi^2(3.96) = 28.09$, $p = .09$) shows that the predictive value of the CBCA scores increases with age. To examine this interaction further, values of the age variable were placed into four approximately equal sized bins, and separate GAMs were run on each. The resulting ogives for these are shown in Figure 3. Simple monotonic curves appear to represent the relationship between CBCA and truthfulness for the older people, but not for the younger groups. It appears either that the relationship between CBCA and truthfulness is different for the age groups, or that the CBCA is only diagnostic of truthfulness above about 16 or 17 points (which the older people do not score below for either true or false statements). Given that these are data created for illustration, it is not appropriate to speculate further about either explanation.

GAMs are useful generalizations of the basic regression models. Like GLMs, they allow different link functions and distributions that are appropriate

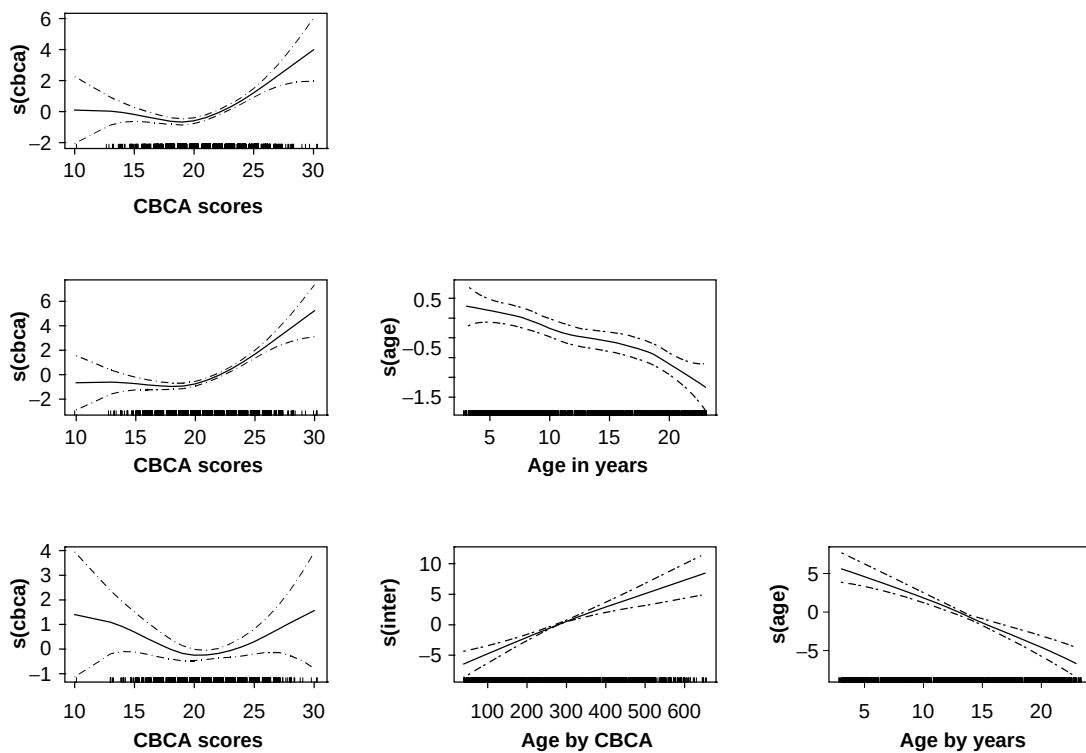


Figure 2 Plots for the Predictor Variables for Three Models Predicting the Probability of a Statement Being Truthful Based on CBCA Score and Age

Notes: The first (upper left hand corner) uses just CBCA score. The second model (row two) also uses age. The third model (row three) also includes the interaction. The dashed lines are for standard errors and rugplots are used to show the univariate distributions.

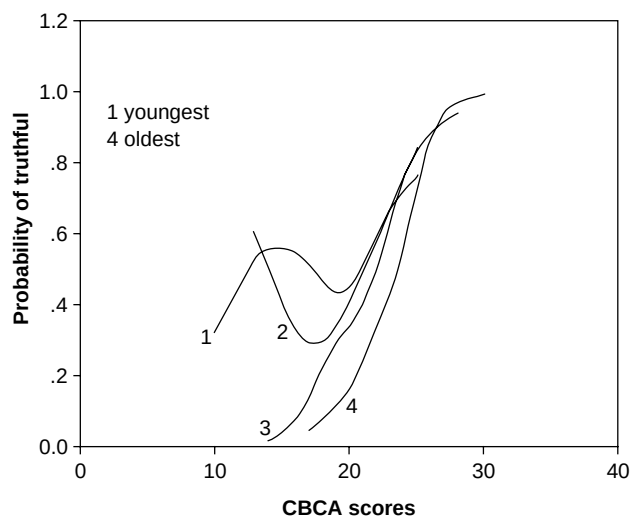


Figure 3 Individual GAMs for Four Different Age Groups

Notes: 1 = 3.0 to 7.9 years, 2 = 7.9 to 13.0 years, 3 = 13.0 years to 18.1 years, and 4 = 18.1 years and higher. There are relatively smooth monotonic curves for the two older age groups. However, the curves for the younger age groups appear more complex, particularly for low CBCA values.

for a large amount of data collected in science. Furthermore, the additive components allow an extremely flexible approach to data modeling. There are several extensions to GAMs not discussed here, such as model selection and regularization techniques, multilevel GAMs, and different types of estimation. Current software allows many different types of curves to be fit within GAMs. Two were illustrative, a theory-driven example where a linear model was compared with a piecewise linear model, and a data-driven example that included an interaction. As algorithms and software advance, these models should become more flexible and more widely used.

—Daniel B. Wright

See also Ogive; Smoothing

Further Reading

Berndt, E. R. (1991). *The practice of econometrics*. New York: Addison-Wesley.

- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
- Vrij, A. (2005). Criteria-based content analysis—A qualitative review of the first 37 studies. *Psychology, Public Policy, & Law*, 11, 3–41.

GENERALIZED METHOD OF MOMENTS

When information on a set of parameters is given in the form of moments (expectations), equations containing this information are called the *moment conditions*. For example, if $y_t = x_t'\theta + u_t$ is considered and the statistician knows a priori that x_t and u_t are uncorrelated, then the corresponding moment conditions are $Ex_t(y_t - x_t'\theta) = 0$. Alternatively, if it is believed that z_t and u_t are uncorrelated for some random variables z_t , then the moment conditions would be $Ez_t(y_t - x_t'\theta) = 0$. In the above examples, the functions $x_t(y_t - x_t'\theta)$ and $z_t(y_t - x_t'\theta)$, whose expectations are set to zero, are called the *moment functions*. In general, for some functions $g(X_t, \theta)$ of random variables X_t and unknown parameter vector θ , the moment conditions are expressed as $Eg(X_t, \theta) = 0$.

Identification and Overidentification

For a given set of moment functions $g(X_t, \theta)$, the true parameter sets the expected moment functions to zero by definition. When $Eg(X_t, \theta) = 0$ at only the true parameter, we say that the true parameter vector is *identified* by the moment conditions. A necessary condition for the identification of the true parameter is that the number of moment conditions should be at least as large as the number of parameters. When the number of moment conditions is exactly equal to the number of parameters (and when the true parameter is identified), we say that the true parameter is *exactly identified*. On the other hand, if there are more moment conditions than necessary, we say that the true parameter is *overidentified*.

Generalized Method of Moments

When there is a set of moment conditions that exactly identifies a parameter vector, *method of moments estimation* is widely used. As the true parameter sets the population moments to zero, the method of moments estimator sets the sample moments to zero. More precisely, when the true parameter is exactly identified by $Eg(X_t, \theta) = 0$, the method of moments estimator $\hat{\theta}$ satisfies $T^{-1}\sum_{t=1}^T g(X_t, \hat{\theta}) = 0$.

If the true parameter is overidentified, that is, if there are more moment conditions than are necessary to identify θ , then it is usually impossible to set the sample moment vector to zero (because there are more equations than parameters). The *generalized method of moments* (GMM) was introduced by Lars Peter Hansen in 1982 in order to handle this case. Let $\bar{g}(\theta) = T^{-1}\sum_{t=1}^T g(X_t, \theta)$ for notational simplicity. Instead of setting the sample moment functions simultaneously to zero (which is usually impossible), Hansen proposed to minimize the quadratic distance of the sample moment vector from zero, that is, to minimize $\bar{g}(\theta)'\bar{g}(\theta)$ with respect to θ over the parameter space. The minimizer is called the *generalized method of moments (GMM) estimator*.

The GMM estimator is consistent and asymptotically normal. In addition, the GMM procedure contains method of moments estimation as a special case. The method of moments estimator sets $\bar{g}(\hat{\theta}) = 0$, in which case the criterion function $\bar{g}(\theta)'\bar{g}(\theta)$ attains the minimal value zero at $\theta = \hat{\theta}$.

Weighted GMM and Optimal GMM

A symmetric and positive definite constant matrix W can be used in the criterion function to form a weighted criterion function $\bar{g}(\theta)'W\bar{g}(\theta)$, whose minimizer is called the *weighted GMM estimator* using the matrix W as weights. Because any symmetric and positive definite matrix can be decomposed into $A'A$ for some nonsingular matrix A (e.g., by a Cholesky decomposition), we observe that any weighted criterion function can be regarded as the (unweighted) quadratic distance of the transformed sample moment vector $A\bar{g}(\theta) = T^{-1}\sum_{t=1}^T Ag(X_t, \theta)$ from

zero. Because W is a constant matrix, A is also a constant matrix, and therefore the transformed moment conditions are also valid, because $E[Ag(X_p, \theta)] = AEg(X_p, \theta) = 0$. This obvious fact shows that any weighted GMM estimator (using a symmetric and positive definite weight matrix) is also consistent and asymptotically normal.

When the moment conditions exactly identify a parameter, a weighted GMM can still be considered. But in that case, the resulting weighted GMM estimator is always equal to the method of moments estimator satisfying $\bar{g}(\hat{\theta}) = 0$. Therefore, weighted GMM needs to be used only in the case of overidentification.

The asymptotic variance of a weighted GMM estimator depends on the associated weight matrix. Optimal weights would correspond to the transformation of the moment conditions such that all the moment functions have identical variance and are pairwise uncorrelated. More specifically, the optimal weight, which yields the most efficient GMM estimator in the class of weighted GMM estimators, is Ω^{-1} , where $\Omega = E[g(X_p, \theta)g(X_p, \theta)']$. The weighted GMM estimator using this optimal weight is called the *optimal GMM estimator*. The optimal GMM estimator is again obviously consistent because $E[Ag(X_p, \theta)] = AEg(X_p, \theta)$, where $A'A = \Omega^{-1}$.

Under the assumption that the random variables X_t are *iid* for all t , and some other technical assumptions, the asymptotic distribution of the optimal GMM estimator is

$$\sqrt{T}(\hat{\theta} - \theta) \rightarrow_d N(0, (D'\Omega^{-1}D)^{-1}),$$

where $D = E\partial g(X_p, \theta)/\partial \theta'$, and both D and Ω are evaluated at the true parameter. This optimal GMM estimator is efficient in the class of consistent estimators based on a given set of moment conditions.

In practice, the optimal weight matrix Ω^{-1} is unknown, and as such, the estimator based on $\bar{g}(\theta)'\Omega^{-1}\bar{g}(\theta)$ is *infeasible*. We can make this procedure *feasible* using a consistent estimate of Ω . Usually, a two-step procedure is used by practitioners.

1. A consistent estimator of Ω is found. Often, this is done using an unweighted GMM, or method of moments estimation with an exactly identifying subset of the moment conditions. If $\tilde{\theta}$ is the resulting estimator, then Ω is estimated by

$$\tilde{\Omega} = T^{-1} \sum_{t=1}^T g(X_p, \tilde{\theta}) g(X_p, \tilde{\theta})'.$$

2. We then minimize $\bar{g}(\theta)'\tilde{\Omega}^{-1}\bar{g}(\theta)$ to yield an efficient GMM estimator.

This estimator is called the *two-step efficient GMM estimator*. The asymptotic distribution of the two-step efficient GMM estimator is identical to that of the (infeasible) optimal GMM estimator using Ω^{-1} as weights.

Sometimes, estimation of Ω and the efficient GMM estimator are repeated until the first step estimator and the second step estimator coincide. This estimator, called the *continuous updating estimator*, also has the same asymptotic distribution as the optimal GMM estimator.

The continuous updating estimator is also obtained by minimizing

$$\bar{g}(\theta)' \left[\frac{1}{T} \sum_{t=1}^T g(X_t, \theta) g(X_t, \theta)' \right]^{-1} \bar{g}(\theta)$$

with respect to θ . The difference between this criterion function and the loss function for the two-step efficient GMM is that in the continuous updating estimation, the weighting matrix is a function of the parameter and is adjusted to attain the global minimum.

When $\hat{\theta}$ is the optimal GMM estimator (feasible or infeasible), the variance-covariance matrix $(D'\Omega^{-1}D)^{-1}$ is estimated by $(\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1}$, where

$$\hat{D} = \frac{1}{T} \sum_{t=1}^T \frac{\partial g(X_t, \hat{\theta})}{\partial \theta'}$$

$$\hat{\Omega} = \frac{1}{T} \sum_{t=1}^T g(X_t, \hat{\theta}) g(X_t, \hat{\theta})'.$$

When computing $\hat{D}$ if the first derivative is not algebraically obtained, we may use a numerical procedure to differentiate the moment functions.

Inferences on the parameters may still be done using this numerical approximation.

When an overidentifying set of moment conditions is available, method of moments estimation using an exactly identifying subset of the moment conditions is always an available option that yields a consistent and asymptotically normal estimator. Moreover, this method of moments estimator based on a subset does not involve weighting and therefore is simpler computationally and conceptually. However, mathematical theorems show that the asymptotic variance of the optimal GMM estimator never increases as moment conditions are added to a given set of moment conditions, which is natural because a bigger set of moment conditions means more information, and an optimal utilization of more information should not yield a worse estimator. So, from the perspective of asymptotic efficiency, it is more desirable to make use of all the moment conditions available and consider optimal GMM. However, when the sample size is small, it is known that too many moment conditions are likely to lead to a poor (e.g., biased) estimator, and methods with which to choose an optimal number of moment conditions are being researched actively.

In order for the GMM estimators (unweighted, weighted, optimal, two-step, and continuous updating) to be consistent and asymptotically normal, the moment conditions must clearly identify the true parameter in the sense that the slope of the moment functions is steep enough to clearly separate the true parameter from its neighborhood. If the moment functions fail this condition, we say that the moment functions *weakly identify* the true parameter. In this case, the two-step efficient GMM estimator has severe bias in general. However, available evidence suggests that the continuous updating estimator has little or no bias. In the example of linear structural equations models, the two-stage least squares estimator with weak instruments is biased, and the estimator corresponds to the two-step efficient GMM estimator. In this same setting, the limited information maximum likelihood estimator corresponds to the continuous updating estimator, and these appear not to have any bias. Formalizing the properties of the GMM estimators in the case of weak identification is an active research area.

Overidentification Test

Now consider testing if the specification $Eg(X_i, \theta) = 0$ is correct. If the true parameter is exactly identified, then it is impossible to test this specification because GMM estimation is achieved by setting $\bar{g}(\hat{\theta}) = 0$. But if the moment conditions overidentify the true parameter, then a test is available based on $\bar{g}(\hat{\theta})$. This test is called the *overidentification test*.

If $H_0: Eg(X_i, \theta) = 0$ is true, then the sample moments $\bar{g}(\hat{\theta})$, evaluated at the GMM (consistent) estimator, will be close to zero, whereas if H_0 is incorrect, then the sample moments will be far from zero. When H_0 is correct, it can be shown that $T\bar{g}(\hat{\theta})'\hat{\Omega}^{-1}\bar{g}(\hat{\theta})$ is approximately χ^2 distributed with degrees of freedom equal to the degree of overidentification, that is, the number of moment conditions minus the number of parameters. If the test statistic is large, then this implies that there are some moment conditions that are not compatible with the others, and so some elements of $Eg(X_i, \theta)$ are not zero, leading us to reject H_0 . Unfortunately, the test does not indicate which of the moment conditions are correctly specified and which are incorrectly specified.

Examples

Any estimator that uses information given in the form of moment conditions can be classified as a GMM estimator. For example, the ordinary least squares estimator is the GMM (or more exactly, the method of moments) estimator using the moment conditions that each regressor is uncorrelated with the error term. The two-stage least squares estimator is another example of GMM based on the moment conditions that the instruments and the error term are uncorrelated. The maximum likelihood estimator can also be regarded as a GMM estimator, because the estimator satisfying the first order condition $T^{-1}\sum_{i=1}^T \partial \log L_i(\hat{\theta})/\partial \theta = 0$ can be regarded as a method of moments estimator using $E \partial \log L_i(\theta)/\partial \theta = 0$ as moment conditions.

Application

Consider a random sample of T observations: $X_1, \dots, X_T$. If we identify the true parameter θ as

$\theta = E(X_t)$, that is, $E(X_t - \theta) = 0$, then the method of moments estimator for θ is $\hat{\theta}_{mm} = T^{-1} \sum_{t=1}^T X_t$, whose variance may be estimated by $\hat{V}_{mm} = T^{-2} \sum_{t=1}^T (X_t - \hat{\theta}_{mm})^2$.

Now suppose that the statistician also knows a priori that the third central moment of X_t is zero, that is, that $E[(X_t - \theta)^3] = 0$. Then we have two moment functions

$$g(X_t, \theta) = \begin{bmatrix} X_t - \theta \\ (X_t - \theta)^3 \end{bmatrix}$$

for the θ parameter. For the two-step efficient GMM estimation, we first get a consistent estimate of θ to construct a consistent estimate for the optimal weighting matrix. We may use the above $\hat{\theta}_{mm}$ to get $\hat{\Omega} = (1/T) \sum_{t=1}^T g(X_t, \hat{\theta}) g(X_t, \hat{\theta})'$, and then the feasible optimal weight is $\hat{\Omega}^{-1}$. Now the two-step efficient GMM estimator minimizes $\bar{g}(\theta)' \hat{\Omega}^{-1} \bar{g}(\theta)$, where $\bar{g}(\theta) = (1/T) \sum_{t=1}^T g(X_t, \theta)$. Let $\hat{\theta}_{gmm}$ denote this two-step efficient GMM estimator. The variance of the asymptotic distribution of $\sqrt{T}(\hat{\theta}_{gmm} - \theta)$ is $(D' \Omega^{-1} D)^{-1}$, where

Table 1 Data: 40 Observations

-0.012	0.442	0.508	1.301	-0.462	0.214	1.693	-0.620
1.281	1.705	1.029	0.807	2.436	-0.349	2.275	2.449
1.593	-0.102	0.756	1.506	0.500	1.407	-0.193	1.095
2.219	1.547	-0.090	2.219	2.003	2.688	0.190	-0.269
1.677	0.576	1.842	-0.107	-0.736	2.318	1.704	1.881

$$D = E \frac{\partial g(X_t, \theta)}{\partial \theta} = - \begin{bmatrix} 1 \\ 3E[(X_t - \theta)^2] \end{bmatrix}$$

and $\Omega = E[g(X_t, \theta) g(X_t, \theta)']$ as before. These D and Ω are estimated consistently by replacing θ with $\hat{\theta}_{gmm}$ and the expectation operator with the sample mean over the T observations. The continuous updating estimation is straightforward.

Table 1 contains a sample of 40 observations of X_t . The method of moments estimator $\hat{\theta}_{mm}$, which is the sample mean of $X_1, \dots, X_{40}$, equals 1.0230 with standard error $\hat{V}^{1/2} = 0.1554$. The S-plus program for the two-step efficient GMM estimation is listed in Table 2, and that for the continuous updating estimation is in Table 3. The resulting two-step GMM estimate is

Table 2 S-Plus and R Code for Two-Step Efficient GMM

```
x <- read.csv ("datafile.csv")
n <- NROW(x)
g.matrix <- function(p) {
  m1 <- x-p
  as.matrix(cbind(m1,m1^3))
}
gmm.func <- function(p) {
  gbar <- as.numeric(colMeans(g.matrix(p)))
  t(gbar)%*%W%*%gbar
}
std.error <- function(p) {
  Omega <- crossprod (g.matrix (p))/n
  D <- -c(1,3*mean((x-p)^2)) # algebraic differentiation
  1/(t(D)%*%solve(Omega)%*%D)/sqrt(n)
}
est1 <- mean(x) # First step MM
W <- solve(crossprod (g.matrix(est1))/n) # Weighting matrix
gmm2 <- nlm(gmm.func,est1) # Second step efficient GMM
est2 <- gmm2$estimate
## Overidentification test
overid <- n*gmm2$minimum
```

Table 3 S-Plus and R Code for Continuous Updating Estimation

```

code continued from Table 2
cue.func <- function(p) {
  gbar <- as.numeric(colMeans (g.matrix (p)))
  W <- solve(crossprod (g.matrix(p))/n)
  t(gbar)%*%W%*%gbar
}
est3 <- nlm(cue.func, est1)$estimate
se3 <- std.error(est3)

```

0.9518 with standard error 0.0456, and the continuous updating estimate is 0.9513 with standard error 0.0456. The sample mean is the method of moments estimator based on the first moment condition only, and the other GMM estimators make use of both moment conditions. As might be expected, the two GMM estimators are more efficient than the sample mean (because more information is used). The overidentification test statistic based on the two-step GMM is computed easily by multiplying the sample size by the minimized criterion function, which is approximately distributed χ^2_1 . For the data set above, the test statistic is 0.3040 with a p value of 0.5814. So, the specification that $E(X_t - \theta) = 0$ and $E[(X_t - \theta)^3] = 0$ is regarded as correct.

—Chirok Han and John Randal

See also Instrumental Variables

Further Reading

- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.
- Lee, M. (1996). *Methods of moments and semiparametric econometrics for limited dependent variable models*. New York: Springer.

GENERALIZED PROCRUSTES ANALYSIS

Generalized Procrustes Analysis (GPA) is a method for determining the degree of agreement, or consensus, among data matrices. For instance, consider 20 judges who rate four brands of coffee on 10 attributes

(e.g., bitterness, richness, smoothness). The ratings for each judge can be recorded in a two-dimensional (10×4) matrix. GPA then can be used to determine the extent to which the judges agree in their views of the four brands of coffee. To the extent the 20 judges do not agree, individual differences in the patterns of ratings can also be examined with GPA. At the heart of the analysis is

a *consensus configuration*, which is derived through a process of scaling, rotating, and averaging the original rating matrices. Each judge's ratings can be compared to this consensus configuration, and an overall *consensus proportion* can be computed that indicates the degree of similarity among the judge's views of the four coffees. GPA is also extremely flexible and can accommodate any number of matrices of varying dimensions. Qualitative judgments or quantitative data can be analyzed, and the matrices must be matched on only a single dimension. For instance, each of the 20 judges could rate the four brands of coffee using a different set of attributes as well as a different number of attributes. The flexibility of GPA can also be seen in the variety of studies in which it has been employed. Researchers have examined individuals' perceptions of food, products, medical treatments, genetic engineering, and personality traits using this technique.

Abbreviated Example

Four managers from a department store freely describe and then rate six of their employees on 5-point scales constructed from their individual descriptive adjectives. A high score on the rating scale indicates that a particular adjective is an accurate description of the employee. The data are reported in Table 1.

The goal of GPA is to assess the degree of similarity among the managers' views of the employees. More specifically, the goal is to determine if the patterns, or profiles, of the six employees are similar across the four managers. Because the focus is on the profiles of the employees, the analysis does not require a fixed set of attributes.

Table 1 Managers' Ratings of Six Employees

Manager 1							Manager 2						
	John	Bob	Amy	Jan	Fred	Jill		John	Bob	Amy	Jan	Fred	Jill
Extraverted	1	1	5	4	5	3	Blunt	2	3	1	3	4	1
Sharing	5	5	2	1	2	3	Patient	5	1	4	2	2	3
Motivated	2	2	4	3	4	1	Creative	5	2	3	4	1	3
Funny	1	2	2	2	4	1	Outgoing	2	4	1	4	3	3
Loud	1	2	2	1	4	1							
Manager 3							Manager 4						
Outgoing	2	4	2	4	5	3	Easygoing	2	3	1	3	4	3
Carefree	2	4	3	4	4	2	Outgoing	1	3	1	5	5	1
Generous	3	2	4	2	2	5	Nurturing	3	3	5	5	1	5
Trusting	3	2	4	3	2	4	Calm	5	2	4	1	2	5
Organized	4	3	3	2	3	3	Intelligent	5	3	5	2	3	3
Athletic	3	2	2	2	4	2							

A number of prescaling methods are typically recommended when conducting GPA. As with any scaling method, the investigator must consider the impact of controlling statistical differences in the data. It is well known, for example, that the mean and standard deviation of z scores are equal to zero and one, respectively. Converting any two variables to z scores will thus equate the two variables on their means and standard deviations. With GPA, three scaling methods are recommended: *centering*, *dimensional*, and *isotropic*.

Centering rescales each manager's ratings such that the mean of each attribute is equal to zero. Isotropic scaling "shrinks" or "expands" each manager's ratings to remove individual differences in scale usage. The ratings for those managers who use relatively few scale values will be expanded with a multiplicative constant greater than one, and the ratings for those managers who use relatively more extreme scale values will be shrunk with a multiplicative constant less than one. The isotropic scaling values for the four managers (.75, 1.01, 1.48, and .94, respectively) indicate that the overall range in the third manager's ratings was less than the range in the other managers' ratings. In fact, it can be seen in Table 1 that the third manager did not use the full range of scale values (1–5). Finally,

dimensional scaling adjusts for differences in the number of attributes in the managers' matrices. The magnitudes of the original values are uniformly increased or decreased depending on the size of the matrix. In this way, matrices with large numbers of attributes will not spuriously influence the results.

Once the original matrices have been rescaled, GPA works through an iterative algorithm in which the six rated employees are maximally aligned across managers. Space does not permit a complete explanation of this algorithm or the equations that underlie GPA, and the Further Reading list at the end of this entry should be consulted. Nonetheless, as the name of the technique implies, Procrustes transformations play an important role in the computations. In essence, the rescaled matrices are sequentially rotated (i.e., transformed) to maximal agreement with an average matrix that is continually updated throughout the process. This average matrix of rescaled ratings is referred to as the *consensus configuration*. At the end of each pass through the four managers' rotated matrices, the consensus configuration is computed and compared to the consensus from the previous iteration. After a number of iterations, changes in the consensus configuration will be negligible, indicating that the procedure has converged on a final solution.

The consensus configuration is therefore the average of rescaled (if scaling options are applied) and rotated ratings, and its dimensionality will equal the largest original matrix. A principal components analysis can be conducted on the consensus configuration, and the employees can be plotted in the space created from the first two components. Results shown in Figure 1 reveal that the managers view John and Amy as highly similar to one another and more similar to Jill than the other three employees.

The attributes generated by the four managers can also be plotted in the component space. It can be seen that Fred, Bob, and Jan are generally viewed as outgoing, easygoing, and carefree compared to Amy, Jill, and John, who are viewed as calm, generous, and patient.

The four managers' individual rating matrices can be compared to the consensus configuration using analysis of variance. The results from such an analysis are shown in Table 2.

As can be seen, an overall *consensus proportion* is produced from the analysis. A value of 1.0 would indicate perfect agreement among the four managers. In such an instance, the four managers' rescaled and rotated matrices would all match the consensus configuration perfectly. Here, the managers' consensus proportion is .79, and a randomization test indicates that this value is statistically significant ($p < .04$).

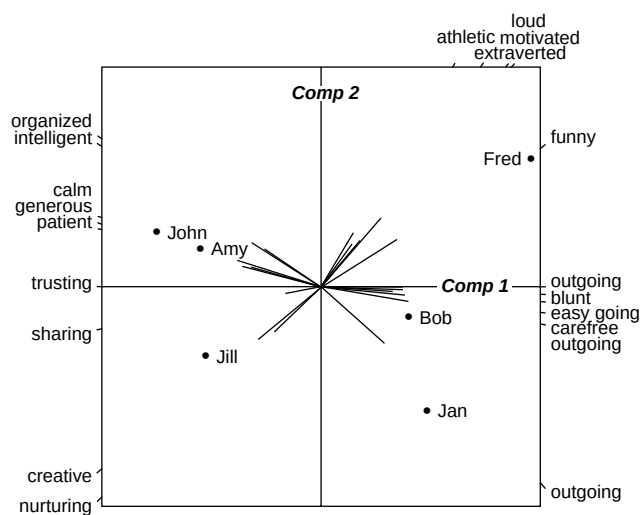


Figure 1 Plot of First and Second Principal Components of the Consensus Grid

Table 2 Analysis of Variance Results

	Consensus	Residual	Total
Employee			
John	16.71	2.33	19.04
Bob	5.47	6.28	11.76
Amy	9.74	5.87	15.61
Jan	11.01	3.55	14.56
Fred	24.10	1.41	25.51
Jill	11.54	1.99	13.52
Grids			
Manager 1		9.87	16.09
Manager 2		5.42	24.61
Manager 3		2.64	29.94
Manager 4		3.51	29.36
Total SS	78.57	21.43	100.00
Consensus Proportion = .79			

Table 2 also shows results for the individual managers and for the four employees. The residual values from these results can be examined to identify points of difference between the individual matrices and the consensus configuration. The residual for the first manager is substantially higher than the others, indicating that this manager is most deviant from the consensus. The residuals for the employees indicate that the ratings for Bob and Amy vary most across the four managers compared to the consensus configuration. The specific residuals shown in Table 3 indicate further that much of the disagreement regarding Bob and Amy is a result of the first manager's ratings. The second manager's ratings of Bob also show relatively high deviance from the consensus configuration.

Table 3 Specific Residuals From Analysis of Variance

	Manager 1	Manager 2	Manager 3	Manager 4
John	0.77	0.75	0.18	0.63
Bob	3.65	1.94	0.62	0.08
Amy	3.39	0.86	0.54	1.07
Jan	0.95	1.31	0.13	1.17
Fred	0.72	0.09	0.21	0.39
Jill	0.38	0.47	0.97	0.17

The first manager's ratings can be submitted to a principal components analysis to examine the nature of the disagreement. The employees and attributes can be plotted in the space created by the first two components, as shown in Figure 2. Comparing the patterns of employees in the consensus configuration in Figure 1 and the first manager's ratings in Figure 2 shows that the manager essentially swapped Bob and Amy. Whereas in the consensus configuration Amy is similar to Jill and John, from the first manager's point of view Bob is similar to Jill and John and Amy is similar to Fred and Jan.

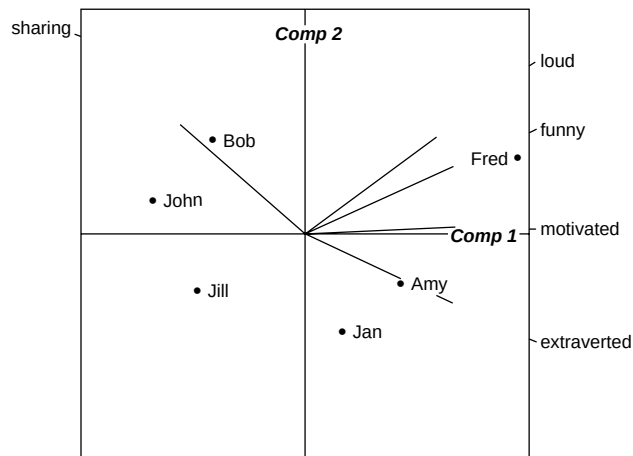


Figure 2 Plot of First and Second Principal Components of the First Manager's Ratings

Additional Issues

The heartbeat of GPA is the consensus configuration. Although its name implies that some sort of agreement has been reached through the analysis, the consensus configuration is essentially a matrix of aggregate values. It is perhaps thus best referred to as the *average configuration* or *centroid configuration*. Nonetheless, the degree of variation around the consensus configuration can be quantified and reported as the consensus proportion. A long-standing criticism of Procrustes transformations is their ability to generate high agreement among even matrices of random numbers. Therefore, the consensus proportion should be tested routinely for statistical significance using a randomization test, as was done above. The issue of prescaling

discussed above is also important. Although all three scaling methods are typically recommended, only the investigator can determine if variability in scale means, scale extremities, or matrix dimensionality represents spurious differences that must be removed from his or her data. Given the convenience of modern computers, conducting the GPA with and without the different scaling options is therefore recommended in order to assess their impact on the results.

—James W. Grice

Further Reading

- Dijksterhuis, G. B., & Gower, J. C. (1991/2). The interpretation of generalized Procrustes analysis and allied methods. *Food Quality and Preference*, 3, 67–87.
- Fewer, L. J., Howard, C., & Shepherd, R. (1997). Public concerns in the United Kingdom about general and specific applications of genetic engineering: Risk, benefit, and ethics. *Science, Technology, & Human Values*, 22, 98–124.
- Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40, 33–51.

GENERALIZED ESTIMATING EQUATIONS

Correlated data sets arise from repeated measures studies where multiple observations are collected from a specific sampling unit (a specific patient's status over time), or from grouped or clustered data where observations are grouped based on sharing some common characteristic (animals in a specific litter). When measurements are collected over time, the term *longitudinal* or *panel data* is preferred. Generalized estimating equations (GEEs) provide a framework for analyzing correlated data. This framework extends the generalized linear models methodology, which assumes independent data. We discuss the estimation of model parameters and associated variances via generalized estimating equation methodology.

The usual practice in model construction is the specification of the systematic and random components of variation. Classical maximum likelihood models then rely on the validity of the specified

components. Model construction proceeds from the (components of variation) specification to a likelihood and, ultimately, an estimating equation. The estimating equation for maximum likelihood estimation is obtained by equating zero to the derivative of the log-likelihood with respect to the parameters of interest. Point estimates of unknown parameters are obtained by solving the estimating equation.

Generalized Linear Models

The theory and an algorithm appropriate for obtaining maximum likelihood estimates where the response follows a distribution in the exponential family was introduced in 1972 by Nelder and Wedderburn. They introduced the term *generalized linear model* (GLM) to refer to a class of models that could be analyzed by a single algorithm. The theoretical and practical application of GLMs has since received attention in many articles and books.

GLMs encompass a wide range of commonly used models such as linear regression, logistic regression for binary outcomes, and Poisson regression for count data outcomes. The specification of a particular GLM requires a link function that characterizes the relationship of the mean response to a vector of covariates. In addition, a GLM requires specification of a variance function that relates the variance of the outcomes as a function of the mean.

The derivation of the iteratively reweighted least squares (IRLS) algorithm appropriate for fitting GLMs begins with the likelihood specification for the exponential family. Within an iterative algorithm, an updated estimate of the coefficient vector may be obtained via weighted ordinary least squares where the weights are related to the link and variance specifications. The estimation is then iterated to convergence where convergence may be defined, for example, as the change in the estimated coefficient vector being smaller than some tolerance.

For any response that follows a member of the exponential family of distributions, $f(y) = \exp\{[y\theta - b(\theta)]/\phi + c(y, \phi)\}$, where θ is the canonical parameter and ϕ is a proportionality constant, we can obtain maximum likelihood estimates of the $p \times 1$ regression

coefficient vector β by solving the estimating equation given by

$$\Psi(\beta) = \sum_{i=1}^n \Psi_i = \sum_{i=1}^n X_i^T (y_i - \mu_i) [\phi V(\mu_i)]^{-1} [\partial \mu_i / \partial \eta_i] = 0_{(p \times 1)}.$$

In the estimation equation, X_i is the i th row of an $n \times p$ matrix of covariates X , $\mu_i = g(x_i\beta)$ represents the expected outcome $E(y) = b'(\theta)$ in terms of a transformation of the linear predictor $\eta_i = x_i\beta$ via a monotonic (invertible) link function $g()$, and the variance $V(\mu_i)$ is a function of the expected value proportional to the variance of the outcome $V(y_i) = \phi V(\mu_i)$. The estimating equation is also known as the score equation because it equates the score vector $\Psi(\beta)$ to zero.

Modelers are free to choose a link function as well as a variance function. If the link-variance pair of functions is chosen from a common member of the exponential family of distributions, the resulting estimates are equivalent to maximum likelihood estimates. However, modelers are not limited to these choices. When one selects variance and link functions that do not coincide to a particular exponential family member distribution, the estimating equation is said to imply existence of a quasi-likelihood, and the resulting estimates are referred to as maximum quasi-likelihood estimates.

The link function that equates the canonical parameter θ with the linear predictor $\eta_i = x_i\beta$ is called the canonical link. If this link is selected, an advantage to interpretation of results is that the estimating equation simplifies to

$$\Psi(\beta) = \sum_{i=1}^n \Psi_i = \sum_{i=1}^n X_i^T (y_i - \mu_i) / \phi = 0_{(p \times 1)}.$$

A second advantage of the canonical link over other link functions is that the expected Hessian matrix is equal to the observed Hessian matrix.

The Independence Model

A basic individual-level model is written in terms of the n individual observations y_i for $i = 1, \dots, n$. When observations may be clustered, due to repeated observations on the sampling unit or because the observations are related to some cluster identifier variable,

the model may be written in terms of the observations y_{it} for the clusters $i = 1, \dots, n$ and the within-cluster repeated, or related, observations $t = 1, \dots, n_i$. The total number of observations is then $N = \sum_i n_i$. The clusters may also be referred to as panels, subjects, or groups. In this presentation, the clusters i are independent, but the within-clusters observations it may be correlated. An independence model, however, assumes that the within-cluster observations are not correlated.

The independence model is a special case of more sophisticated correlated data approaches (such as GEE). This model assumes that there is no correlation within clusters. Therefore, the model specification is in terms of the individual observations y_{it} . Although the independence model assumes that the repeated measures are independent, the model still provides consistent estimators in the presence of correlated data. Of course, this approach is paid for through inefficiency, although the efficiency loss is not always large. As such, this model remains an attractive alternative because of its computational simplicity. The independence model also serves as a reference model in the derivation of diagnostics for more sophisticated models for clustered data (such as GEE models).

Analysts can use the independence model to obtain point estimates along with standard errors based on the modified sandwich variance estimator to ensure that inference is robust to any type of within-cluster correlation. Although the inference regarding marginal effects is valid (assuming that the model for the mean is correctly specified), the estimator from the independence model is not efficient when the data are correlated.

The validity of the (naive) model-based variance estimators depends on the correct specification of the variance; in turn, this depends on the correct specification of the working correlation model. A formal justification for an alternative estimator known as the *sandwich variance estimator* is given in Huber.

It should be noted that assuming independence is not always conservative; the model-based (naive) variance estimates based on the observed or expected Hessian matrix are not always smaller than those of the modified sandwich variance estimator. Because

the sandwich variance estimator is sometimes called the *robust variance estimator*, this result may seem counterintuitive. However, it is easily seen by assuming negative within-cluster correlation leading to clusters with both positive and negative residuals. The clusterwise sums of those residuals will be small, and the resulting modified sandwich variance estimator will yield smaller standard errors than the model-based Hessian variance estimators.

Other obvious approaches to the nested structure assumed for the data include fixed effects and random effects models. Fixed effects models incorporate a fixed increment to the model for each group, whereas random effects models assume that the incremental effects from the groups are from a common random distribution; in such a model, the parameters of the assumed random effects distribution are estimated rather than the effects. In the example at the end of this entry, we consider two different distributions for random effects in a Poisson model.

Subject-Specific Versus Population-Averaged Models

There are two main approaches to dealing with correlation in repeated or longitudinal data. One approach focuses on the marginal effects averaged across the individuals (population-averaged approach), and the second approach focuses on the effects for given values of the random effects by fitting parameters of the assumed random effects distribution (subject-specific approach).

The population-averaged approach models the average response for observations sharing the same covariates (across all of the clusters or subjects), whereas the subject-specific approach explicitly models the source of heterogeneity so that the fitted regression coefficients have an interpretation in terms of the individuals.

The most commonly described GEE model is a population-averaged approach. Although it is possible to derive subject-specific GEE models, such models are not currently supported in commercial software packages and so do not appear nearly as often in the literature.

The basic idea behind this approach is illustrated as follows. We consider the estimating equation for GLMs; that estimating equation, in matrix form, is for the exponential family of distributions

$$\begin{aligned}\Psi(\beta) &= \sum_{i=1}^n \Psi_i \\ &= \sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}[\partial \mu_i / \partial \eta_i] \mathbf{V}^{-1}(\mu_i)(y_i - \mu_i) / \phi \\ &= \sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}[\partial \mu_i / \partial \eta_i] \mathbf{V}^{-1/2}(\mu_i) \\ &\quad \mathbf{I}_{(n \times n)} \mathbf{V}^{-1/2}(\mu_i)(y_i - \mu_i) / \phi = \mathbf{0}_{(p \times 1)}.\end{aligned}$$

Assuming independence, $\mathbf{V}^{-1}(\mu_i)$ is clearly an $n_i \times n_i$ diagonal matrix that can be factored with an identity matrix in the center playing the role of the correlation of observations within a given group or cluster. This corresponds to the independence model already mentioned.

The genesis of the original population-averaged generalized estimating equations is to replace the identity matrix with a parameterized working correlation matrix $\mathbf{R}(\alpha)$. To address correlated data, the working correlation matrix imposes structural constraints. In this way, the independence model is a special case of the GEE specifications where $\mathbf{R}(\alpha)$ is an identity matrix.

Formally, Liang and Zeger introduce a second estimating equation for the structural parameters of the working correlation matrix. The authors then establish the properties of the estimators resulting from the solution of these estimating equations. The GEE moniker was applied because the model is derived through a generalization of the GLM estimating equation; the second-order variance components are introduced directly into the estimating equation rather than appearing in consideration of a multivariate likelihood.

Several software packages support estimation of these models. These packages include R, SAS, S-PLUS, Stata, and SUDAAN. R and S-PLUS users can easily find user-written software tools for fitting

GEE models, whereas such support is included in the other packages.

Estimating the Working Correlation Matrix

One should consider carefully the parameterization of the working correlation matrix, because including the correct parameterization leads to more efficient estimates. We want to consider this choice carefully even if we employ the modified sandwich variance estimator in the calculation of standard errors and confidence intervals for the regression parameters. Although the use of the modified sandwich variance estimator ensures robustness in the case of misspecification of the working correlation matrix, the advantage of more efficient point estimates is still worth this effort. There is no controversy as to the fact that the GEE estimates are consistent, but there is some controversy as to how efficient they are. This controversy centers on how well the correlation parameters can be estimated.

Typically, a careful analyst chooses some small number of candidate parameterizations. Pan also discusses the quasi-likelihood information criterion (QIC) measures for choosing between candidate parameterizations. This criterion measure is similar to the well-known Akaike information criterion (AIC).

The most common choices for the working correlation $\mathbf{R}$ matrix are given by parameterizing the elements of the matrix, as shown in Table 1.

The independence model admits no extra parameters, and the resulting model is equivalent to a generalized linear model specification, and the

Table 1 Common Correlation Structures

Independent	$R_{uv} = 0$
Exchangeable	$R_{uv} = \alpha$
Autocorrelated – AR(1)	$R_{uv} = \alpha^{u-v}$
Stationary (k)	$R_{uv} \alpha_{ u-v }$ if $ u - v \leq k$ 0 otherwise
Nonstationary (k)	$R_{uv} \alpha_{(u,v)}$ if $ u - v \leq k$ 0 otherwise
Unstructured	$R_{uv} = \alpha_{(u,v)}$

Note: Values are given for $u \neq v$; $R_{uu} = 1$.

exchangeable correlation parameterization admits one extra parameter. The most general approach is to consider the unstructured (only imposing symmetry) working correlation parameterization, which admits $M(M - 1)/2 - M$ extra parameters, where $M = \max_i \{n_i\}$. The *exchangeable* correlation specification is also known as *equal correlation*, *common correlation*, and *compound symmetry*.

The elements of the working correlation matrix are estimated using the Pearson. Estimation alternates between estimating the regression parameters β , assuming the current estimates of α are true, and then

assuming β estimates are true to obtain residuals to update the estimate of α .

Example

To highlight the interpretation of GEE analyses and point out the alternate models, we focus on a simple example (Table 2).

These data have been analyzed in many forums and are from a panel study on Progabide treatment of epilepsy. Baseline measures of the number of seizures in an 8-week period were collected and recorded as

Table 2 Data From Progabide Study on Epilepsy
(59 Patients Over 5 Weeks)

<i>id</i>	<i>age</i>	<i>trt</i>	<i>base</i>	<i>s1</i>	<i>s2</i>	<i>s3</i>	<i>s4</i>
1	31	0	11	5	3	3	3
2	30	0	11	3	5	3	3
3	25	0	6	2	4	0	5
4	36	0	8	4	4	1	4
5	22	0	66	7	18	9	21
6	29	0	27	5	2	8	7
7	31	0	12	6	4	0	2
8	42	0	52	40	20	23	12
9	37	0	23	5	6	6	5
10	28	0	10	14	13	6	0
11	36	0	51	26	12	6	22
12	24	0	33	12	6	8	5
13	23	0	18	4	4	6	2
14	36	0	42	7	9	12	14
15	26	0	87	16	24	10	9
16	26	0	50	11	0	0	5
17	28	0	18	0	0	3	3
18	31	0	111	37	29	28	29
19	32	0	18	3	5	2	5
20	21	0	20	3	0	6	7
21	29	0	12	3	4	3	4
22	21	0	9	3	4	3	4
23	32	0	17	2	3	3	5
24	25	0	28	8	12	2	8
25	30	0	55	18	24	76	25
26	40	0	9	2	1	2	1
27	19	0	1-	3	1	4	2
28	22	0	47	13	15	13	12
29	18	1	76	11	14	9	8
30	32	1	38	8	7	9	4
31	20	1	19	0	4	3	0
32	20	1	10	3	6	1	3
33	18	1	19	2	6	7	4
34	24	1	24	4	3	1	3
35	30	1	31	22	17	19	16
36	35	1	14	5	4	7	4
37	57	1	11	2	4	0	4
38	20	1	67	3	7	7	7
39	22	1	41	4	18	2	5
40	28	1	7	2	1	1	0
41	23	1	22	0	2	4	0
42	40	1	13	5	4	0	3
43	43	1	46	11	14	25	15
44	21	1	36	10	5	3	8
45	35	1	38	19	7	6	7
46	25	1	7	1	1	2	4
47	26	1	36	6	10	8	8
48	25	1	11	2	1	0	0
49	22	1	151	102	65	72	63
50	32	1	22	4	3	2	4
51	25	1	42	8	6	5	7
52	35	1	32	1	3	1	5
53	21	1	56	18	11	28	13
54	41	1	24	6	3	4	0
55	32	1	16	3	5	4	3
56	26	1	22	1	23	19	8
57	21	1	25	2	3	0	1
58	36	1	13	0	0	0	0
59	37	1	12	1	4	3	2

Table 3 Estimated Incidence Rate Ratios and Standard Errors for Various Poisson Models

<i>Model</i>	<i>time</i>	<i>trt</i>	<i>age</i>	<i>baseline</i>
Independence	0.944 (0.019, 0.033)	0.832 (0.039, 0.143)	1.019 (0.003, 0.010)	1.095 (0.002, 0.006)
Gamma RE	0.944 (0.019)	0.810 (0.124)	1.013 (0.011)	1.116 (0.015)
Gaussian RE	0.944 (0.019, 0.033)	0.760 (0.117, 0.117)	1.011 (0.011, 0.009)	1.115 (0.012, 0.011)
GEE(exch)	0.939 (0.019, 0.019)	0.834 (0.058, 0.141)	1.019 (0.005, 0.010)	1.095 (0.003, 0.006)
GEE(ar 1)	0.939 (0.019, 0.019)	0.818 (0.054, 0.054)	1.021 (0.005, 0.003)	1.097 (0.003, 0.003)
GEE(unst)	0.951 (0.017, 0.041)	0.832 (0.055, 0.108)	1.019 (0.005, 0.009)	1.095 (0.003, 0.005)

base for 59 patients. Four follow-up 2-week periods also counted the number of seizures; these were recorded as *s1*, *s2*, *s3*, and *s4*. The *base* variable was divided by four in our analyses to put it on the same scale as the follow-up counts. The *age* variable records the patient's age in years, and the *trt* variable indicates whether the patient received the Progabide treatment (value recorded as one) or was part of the control group (value recorded as zero).

An obvious approach to analyzing the data is to hypothesize a Poisson model for the number of seizures. Because we have repeated measures, we can choose a number of alternative approaches. In our illustrations of these alternative models, we use the baseline measure as a covariate along with the *time* and *age* variables.

Table 3 contains the results of several analyses. For each covariate, we list the estimated incidence rate ratio (exponentiated coefficient). Following the incidence rate ratio estimates, we list the classical and sandwich-based estimated standard errors. We did not calculate sandwich-based standard errors for the gamma distributed random effects model.

We emphasize again that the independence model coupled with standard errors based on the modified sandwich variance estimator is a valid approach to modeling data of this type. The weakness of the approach is that the estimators will not be as efficient as a model including the true underlying within-cluster correlation structure. Another standard

approach to modeling this type of repeated measures is to hypothesize that the correlations are due to individual-specific random intercepts. These random effects (one could also hypothesize fixed effects) will lead to alternate models for the data.

Results from two different random effects models are included in the table. The gamma-distributed random effects model is rather easy to program and fit to data because the log-likelihood of the model is in analytic form. On the other hand, the normally distributed random effects model has a log-likelihood specification that includes an integral. Sophisticated numeric techniques are required for the calculation of such a model.

We could hypothesize that the correlation follows an autoregressive process because the data are collected over time. However, this is not always the best choice in an experiment because we must believe that the hypothesized correlation structure applies to both the treated and untreated groups.

The QIC values for the independence, exchangeable, ar1, and unstructured correlation structures are respectively given by -5826.23 , -5826.25 , -5832.20 , and -5847.91 . This criterion measure indicates a

Table 4 Fitted Correlation Matrices

1.00				1.00			
0.51	1.00			0.25	1.00		
0.26	0.51	1.00		0.42	0.68	1.00	
0.13	0.26	0.51	1.00	0.22	0.28	0.58	1.00

preference for the unstructured model over the autoregressive model. The fitted correlation matrices for these models (printing only the bottom half of the symmetric matrices) are given by Table 4.

—James W. Hardin

Further Reading

- Glonek, G. F. V., & McCullagh, R. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society—Series B*, 57, 533–546.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings for the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–223). Berkeley: University of California Press.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society—Series A*, 135(3), 370–384.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120–125.

GERONTOLOGICAL APPERCEPTION TEST

The Gerontological Apperception Test, developed by Wolk and Wolk in 1971, is a projective test published by Behavioral Publications that is designed to compensate for the reputed weakness of many apperceptive tests in the assessment of older adults. Because at least one older adult is depicted in a situation frequently encountered by the aged, identification with the test stimuli is assumed to be evocative of responses. This enhances an understanding of the aged personality and his or her reactions to common situations.

Specifically, the Gerontological Apperception Test consists of a set of 14 achromatic cards, each card reflecting a situation with which older adults could identify. The pictures are designed to elicit more relevant themes such as isolation, loss of physical

mobility/sexuality, dependency, and ageism. There is no standard set of cards to be administered. Cards are selected for administration based on suspected personal issues and concerns of each individual subject. At the time of the initial publication of the Gerontological Apperception Test, there was not a standard scoring procedure. Protocols were typically analyzed for the presence of themes.

The Gerontological Apperception Test has been criticized on the bases of (a) the negative tone of the pictures (all achromatic), (b) the stereotypic presentation of older people on the cards, and (c) no accepted scoring system that would permit the development of norms to guide clinical use. However, a revised and more differentiated scoring system for the Gerontological Apperception Test was developed by Hayslip and his colleagues.

The literature on the Gerontological Apperception Test is not extensive, nor does the literature demonstrate that the test is more successful than the Thematic Apperception Test in eliciting relevant themes from older adults. Therefore, although the test shows promise for clinical use with older adults, more research is required before the clinical utility of the Gerontological Apperception Test has been demonstrated effectively.

—Paul E. Panek

See also Thematic Apperception Test

Further Reading

- Hayslip, B., Jr., Francis, J., Radika, L. M., Panek, P. E., & Bosmajian, L. (2002). The development of a scoring system for the Gerontological Apperception Test. *Journal of Clinical Psychology*, 58, 471–478.
- Wolk, R. L., & Wolk, R. B. (1971). *The Gerontological Apperception Test*. New York: Behavioral Publications.

Gf-Gc THEORY OF INTELLIGENCE

There are a number of widely known and respected theories of human intelligence (e.g., Howard Gardner's

Theory of Multiple Intelligences, Robert Sternberg's Triarchic Theory of Intelligence), but in recent years, the Gf-Gc theory has become increasingly important in the field of intellectual assessment. This is at least partially due to the fact that the Gf-Gc theory is based on factor analytical studies of the results from IQ tests. Many of the other theories, although intriguing, do not have a means of actually measuring their constructs in an individual.

In the early 1900s, Charles Spearman applied statistical analyses to the concept of mental ability and arrived at the conclusion that there is one general factor (g) that is related to all aspects of intelligence. Although this is expressed in many different ways, underlying it all is one thing that we think of as intelligence. This concept took root and is the basis for the single IQ score that was generated with early measures of intelligence and, although controversial, continues to be reflected in the Full Scale IQ or similar score found on most measures of cognitive ability today.

It was not until the 1940s that Raymond Cattell, building on Spearman's work, proposed the existence of two general types of intelligence: fluid intelligence (Gf) and crystallized intelligence (Gc). Fluid intelligence was related to biological and neurological factors and is exemplified by inductive and deductive reasoning. Although experience may influence it indirectly (e.g., introduction of new paradigms allows for different ways of organizing problems), it is not dependent upon learned information. Conversely, crystallized intelligence was seen as being the direct result of experience, learning, and education, and was relatively free from the influence of biological and neurological factors. This dichotomy was frequently thought of loosely as innate and learned abilities, right and left hemisphere abilities, or nonverbal and verbal abilities, respectively. Although useful conceptually, these alternative ways of thinking about the two types of intelligence were not fully supported by the research and remain controversial at best.

In the early 1980s, John Horn, using the decades of factor analytic research on human cognitive abilities since Cattell's original postulation, added to the original Gf-Gc theory to form what became known as the Cattell-Horn Gf-Gc theory of intelligence. This new

theory contained 9 to 10 broad abilities by the mid-1990s and began to be used more and more as a basis for interpreting the results of intelligence tests.

At this same time, John Carroll conducted a meta-analysis of more than 400 different data sets that had been collected from 1925 on. He looked at the raw scores, conducted exploratory factor analyses, and concluded that the results fit a hierarchical three-stratum model: 69 narrow abilities at the first level; 8 broad abilities (that roughly corresponded to the broad abilities articulated by Horn-Cattell) at the second level; and a general factor, g, above them all at the third level. Subsequently, these two models were merged by Kevin McGrew and others to form the Cattell-Horn-Carroll (CHC) theory of cognitive abilities. CHC theory and Gf-Gc theory are now essentially analogous.

Current articulations of Gf-Gc theory typically include 10 different broad abilities. Two of these are usually found on measures of achievement and are considered basic academic skills. These two are Quantitative Knowledge (Gq), which represents an individual's store of acquired mathematical knowledge, and Reading/Writing Ability (Grw), which represents an individual's acquired store of knowledge related to the comprehension of written material and the expression of thoughts in writing. It is important to differentiate Gq from Quantitative Reasoning (RQ), which is a narrow ability that is part of Fluid Reasoning (Gf). Gq is evident in applying a mathematical formula to arrive at a solution, whereas RQ would be involved in finding the missing number in a number series. Grw is not clearly defined but appears to consist of reading decoding, reading comprehension, reading speed, spelling, grammar and punctuation, and written expression. Both Gq and Grw are generally found on measures of achievement and not on measures of intelligence.

The remaining eight abilities—Fluid Reasoning (Gf); Crystallized Intelligence (often referred to as Comprehension-Knowledge, or Gc); Short-Term Memory (Gsm); Visual Processing (Gv); Auditory Processing (Ga); Long-Term Storage and Retrieval (Glr); Processing Speed (Gs); and Decision/Reaction Time or Speed (Gt)—are typically found on

intelligence tests. Depending upon the measure of intelligence, different abilities may be emphasized or absent. For example, the Wechsler Scales do not have measures of Auditory Processing (Ga) but have multiple measures of Comprehension-Knowledge (Gc), and the Woodcock Johnson-III (WJ-III) has multiple measures of all but Decision/Reaction Time (Gt).

Fluid Reasoning (Gf) is basically novel problem solving and reasoning. It subsumes the narrow abilities of General Sequential Reasoning, Induction, Quantitative Reasoning, Piagetian Reasoning, and Reasoning Speed. In everyday life, it is the ability to form concepts, apply logic, manipulate abstractions and relations, and solve problems that include novel information and procedures. It is particularly important in fields that require the logical application of creativity (e.g., engineering, research science).

Crystallized Intelligence, or Comprehension-Knowledge (Gc), is an individual's breadth and depth of knowledge, the ability to communicate (especially verbally) that knowledge, and the ability to apply that knowledge in reasoning. Under the broad Gc ability, we find the narrow abilities of Language Development; Lexical Knowledge; Listening Ability; Information (General, Cultural, Science, Geography); Communication Ability; Oral Production & Fluency; Grammatical Sensitivity; and Foreign Language (Proficiency and Aptitude). Of all the factors, Gc is the most heavily culturally loaded because it depends upon experience for development and is the most resistant to neurological damage, such as traumatic brain injury. It is also the single factor that is likely to increase as we age and gain additional experience and the factor mostly closely correlated to academic performance.

Short-Term Memory (Gsm) is the ability to apprehend and hold information in immediate awareness and to use it within a few seconds. It is composed of the narrow abilities of Memory Span, Working Memory, and Learning Abilities (also under Glr). Measurement of Gsm is highly sensitive to attentional problems (have to actually perceive it before you can remember it) and anxiety. It is a foundational ability that will influence many other abilities. For example, you cannot solve a multivariable problem if you can keep only one variable in awareness at a time.

Visual Processing (Gv) is the ability to perceive and manipulate visual shapes and to analyze and synthesize visual information. It includes the narrow abilities of Spatial Relations, Visual Memory, Closure Speed, Flexibility of Closure, Spatial Scanning, Serial Perceptual Integration, Length Estimation, Perceptual Illusions, Perceptual Alterations, Visualization, and Imagery. It is the ability to see things in your mind's eye; to manipulate those things mentally (rotate, rearrange, resize, assemble, take apart, etc.); and to use these skills to solve real-world problems. Typically, someone high in Gv will be considered a visual learner.

Auditory Processing (Ga) is the analog of Gv with auditory stimuli rather than visual. It is the ability to perceive, discriminate, analyze, and synthesize auditory information. The narrow abilities under Ga are Phonetic Coding (Analysis and Synthesis), Speech Sound Discrimination, Resistance to Auditory Distortion, Memory for Sound Patterns, General Sound Discrimination, Temporal Tracking, Musical Discrimination and Judgment, Sound-Intensity/Duration Discrimination, Sound-Frequency Discrimination, Hearing and Speech Threshold Factors, Absolute Pitch, and Sound Localization. Although understanding of a language is not necessary for measuring Ga, it is likely to be very important in the development of language and is related to musical ability.

Long-Term Storage and Retrieval (Glr) is the ability to store information in and then fluently retrieve that information from long-term memory through the use of association. It is the process of storage and retrieval, and not the information that is actually stored, that constitutes Glr. Under the umbrella of Glr are the following narrow abilities: Associative; Meaningful and Free Recall Memory; Fluency (Ideational, Associational, and Expressional) Naming Facility; Word and Figural Fluency; Figural Flexibility; Sensitivity to Problems; Originality/Creativity; and Learning Abilities. Clearly, Glr is of critical importance in everyday life to be able to not only acquire and store new information, but to be able to access that information when needed. One analogy that captures the essence of Glr is that of the fisherman's net. The knots or intersections represent the pieces of information and the strands that you must traverse in Glr.

Processing Speed (Gs) is the ability to perform cognitive tasks fluently and automatically, especially when under pressure to maintain focused attention and concentration. It is basically the ability to make routine things automatic so that each step of the task does not have to be processed (e.g., alphabetizing files or other clerical tasks). It consists of Perceptual Speed, Rate-of-Test Taking, Number Facility, and Semantic Processing Speed. Individuals low in processing speed not only may find themselves taking more time to perform tasks, but also may find those tasks more effortful and tiring because they cannot do them automatically.

—Steve Saladin

See also Intelligence Quotient; Intelligence Tests

Further Reading

- Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Boston: Allyn & Bacon.
- Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.
- Sternberg, R. J. (2000). *Handbook of intelligence*. New York: Cambridge University Press.

GOODENOUGH HARRIS DRAWING TEST

The Goodenough Harris Drawing Test, published by Psychological Corporation, is a nonverbal test of mental ability that is appropriate as either a group or an individual test. The test takes 10 to 15 minutes to administer to children ages 3 to 15. The directions given to the children are simple: "Make a picture of a man; make the very best picture you can." The current version of the test is essentially the 1963 revision of the original Draw a Man Test.

The purpose of constructing the Goodenough Harris Drawing Test was to substitute the concept of intelligence with the notion of *intellectual maturity* or, more accurately, *conceptual maturity*. Thus, intellectual maturity means the ability to form concepts of an

abstract character. This encompasses (a) the ability to perceive (i.e., to distinguish between likenesses and differences), (b) the ability to abstract (i.e., to put into groups objects according to likenesses and differences), and (c) the ability to generalize (i.e., to assign newly experienced objects to the correct class). Therefore, evaluation of children's drawings of the human figure helps to measure the complexity of the child's concept formation. The human figure is employed because it is the most familiar and significant figure for the children.

The evaluation of children's drawings is carried out by two different scoring procedures, the Point Scale and the 12-Point Quality Scale. According to the first procedure, each item is rated as pass or fail (1 point or 0), which is based on the presence or absence of a body part or a specific detail (e.g., eyes are present). The Draw a Man Test has 73 items and the Draw a Woman Test 71 items. The scorable items of both drawings are chosen on the basis of (a) age differentiation, (b) relationship to tests of general intelligence, and (c) diversification of children of lower to higher intellectual ability. The score (marking) on the Goodenough Harris Drawing Test is a single one. A detailed scoring guide is offered in the test manual. The second procedure relies on a 12-Point Quality Scale in which 1 indicates the lowest category and 12 the highest.

Norms for both the Point Scale and the 12-Point Quality Scale are provided. The psychometric properties are good, although the test is better employed for children ages 5–15. The test is simple to administer, is enjoyed by the children, and measures general intelligence. However, its cross-cultural use is questionable. The test works well with younger children, especially those of lower intellectual abilities, language handicapped, minority, and bilingual children.

—Demetrios S. Alexopoulos

See also Personality Tests

Further Reading

- Harris, D. D. (1963). *Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough Harris Draw a Man Test*. New York: Harcourt, Brace & World.

Sattler, J. M. (1992). *Assessment of children: Revised and updated* (3rd ed.). San Diego, CA: Author.

Scott, L. D. (1981). Measuring intelligence with the Goodenough-Harris Drawing Test. *Psychological Bulletin*, 89, 483–505.

Goodenough-Harris drawing test: <http://gri.gallaudet.edu/~catraxle/INTELLEC.html#goodenough>

GOODNESS-OF-FIT TESTS

Most of the commonly used statistical methods are known to be parametric tests, which impose distributional assumptions on the data. For instance, the t test is popular in comparing the means of two independent samples. This test assumes that the underlying distribution from which each of the samples came is a normal distribution. This assumption is critical, especially in cases where the sample sizes are small. If this distributional assumption is not met, results may be invalid and misleading. Some statistical procedures based on the normal distribution are still approximately valid regardless of the distribution of the data as long as the sample size is large enough. The problem is that it is not clear what “large enough” means. In some cases, a sample size of 30 is large enough, whereas in other cases, a sample size of 30 is not sufficient. On the other hand, one may opt to use nonparametric statistical methods, which do not assume a specific form of the distribution of the samples. These procedures are valid regardless of the sample size. However, it is a well-known fact that nonparametric tests are not as powerful as parametric tests; that is, a nonparametric test requires a larger sample size than its corresponding parametric test to detect a difference, if one truly exists, as long as the distributional assumption of the parametric test is satisfied. Therefore, it is important, especially when sample sizes are small, that the distributional assumption of parametric tests be checked and validated before reporting the results of the statistical analyses are reported. Goodness-of-fit (GOF) tests provide methods to achieve this purpose.

The null and alternative hypotheses of the GOF tests are as follows:

Null hypothesis (H_0): assumed distribution has a good fit

Alternative hypothesis (H_a): assumed distribution is not a good fit

A GOF test does not try to prove that the underlying distribution is true. Instead, it starts by assuming that the data follow the underlying distribution. It rejects this assumption if there is strong evidence of violation of this assumption, and it does not suggest an alternative distribution to consider. A GOF test does not give any information on how the data deviate from the hypothesized distribution; for this reason, it is highly recommended that GOF tests be accompanied by graphical representation of the data distribution, such as a probability plot, if one exists for the distribution being tested, or a histogram. Moreover, it is possible that GOF tests will not reject a number of distributions, implying that these distributions are a good fit to the data. GOF tests are not designed to choose which among these distributions best fits the data.

Numerous goodness-of-fit tests exist, and they can range from simple to very complex depending on whether the underlying distribution is univariate or multivariate. The most popular and simplest univariate GOF tests are the chi-square goodness-of-fit test and the Kolmogorov-Smirnov test. The chi-square GOF test may be applied whether the underlying distribution is discrete and continuous. The Kolmogorov-Smirnov test applies only when the underlying distribution is continuous. Both of these tests are available in most statistical packages.

Chi-Square Goodness-of-Fit Test

The idea behind the chi-square GOF test is simple. It compares the observed proportion to the expected proportion based on the assumed distribution. Because distributions depend on parameters that are typically unknown, these parameters are first estimated from the data. These estimates will be plugged in to the distribution to compute the expected proportion.

Consider a data set with n observations. The following are steps in a chi-square GOF test.

1. Define K classes in which to assign each observation. If data are continuous (interval scale), the classes defined in constructing a histogram may be used. In this case, classes need not be of equal interval size.

2. Count the number of observations that fall on the i th class and denote this by O_i .

3. Compute the expected number of observations that will fall on the i th class based on the underlying distribution and denote this by E_i . In the continuous case, where the cumulative distribution function is denoted by $F(x)$, the expected number falling in the interval $[L_i, U_i]$ is

$$E_i = n * [\hat{F}(U_i) - \hat{F}(L_i)],$$

where $\hat{F}$ is the cumulative distribution function using the estimated values of any unknown parameters.

4. Compute the test statistic, χ^2 , as

$$\chi^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i}.$$

5. Finally, reject the null hypothesis at an approximate level α if

$$\chi^2 > \chi^2_{\alpha, K-1-p},$$

where K is the number of classes, p is the number of estimated parameters in the underlying distribution, and α is the $(1 - \alpha)$ th percentile of a chi-squared distribution with degrees of freedom $K - 1 - p$.

The chi-square GOF test requires adequate sample size to be valid. Furthermore, it requires all cells to be nonempty and have expected counts of at least 5. Collapsing cells is a common remedy to this problem. However, if cells represent categories that are not related, collapsing may not be a good idea.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) goodness-of-fit test compares the theoretical cumulative distribution

function, $F(x)$, with the empirical distribution function, which is an estimate of the cumulative distribution function based on the data. If the distribution being tested is a good fit, then the theoretical and empirical cumulative distribution functions should be close to each other. The KS statistic is based on the maximum distance between these two functions. As previously mentioned, a major limitation of the KS test is that it cannot be applied to discrete data.

The steps in the KS goodness-of-fit test are as follows.

1. Arrange the observations in increasing order and label them as $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, so that $X_{(1)}$ and $X_{(n)}$ are the smallest and largest observations, respectively.

2. Compute the KS test statistic based on the formula

$$D = \max_{1 \leq i \leq n} \left[F(X_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F(X_{(i)}) \right]$$

3. Finally, if the value of the D statistic is larger than the KS critical value at a given level of significance, then we reject the hypothesis that the assumed distribution is a good fit to the data. Tables of KS critical values are readily available in the literature. These critical values are applicable only when the parameters of the distribution are completely specified. Otherwise, the critical values are computed via simulation.

Other GOF Tests

Other GOF tests are available for continuous data, such as Anderson-Darling, empirical distribution function, and Cramer-von Mises tests, to name a few. GOF tests for discrete data are available for specific discrete distributions, such as the Poisson distribution.

GOF tests are also available in other commonly used statistical procedures. In ordinary regression analysis and analysis of variance, the distribution of the error term is assumed to follow a normal distribution, and hence, the response variable also has a

normal distribution. GOF tests are not applied to the response variable because its distribution depends on the predictors or factors in the model. To test the distributional assumption of these procedures, univariate GOF tests are instead applied to residuals, whose distribution does not depend on the predictor variables or factors in the model.

A cautionary note: GOF tests, just like any statistical test of a hypothesis, will tend to reject the null hypothesis as the sample size increases. Therefore, one should be very careful in interpreting the results of GOF tests when the sample sizes are very large. Sometimes, performing a GOF test is no longer advisable. Instead, graphical techniques are better suited to assessing distributional assumptions.

An Illustration

One hundred observations (Y) were generated using SAS from a normal distribution with mean 3 and variance 0.5. A set of X observations was obtained from Y via the transformation $X = e^Y$. In this case, X is known to have a lognormal distribution. GOF tests were performed on X and Y observations testing the fit of the normal and lognormal distributions.

Figure 1 displays the SAS code used in this illustration, which provides results of Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, and chi-square goodness-of-fit tests. Figure 2 displays the histogram of the X observations with the best fit normal and lognormal densities showing that lognormal is a better fit than normal. Portions of the output from SAS displaying the results of the GOF tests on X observations are presented in Tables 1 and 2. Table 1 contains the result of fitting a normal distribution. The interval size used for the chi-square test is 30 units, with the midpoint starting at 0. Except for the chi-square test, all other tests are significant at the 5% level of significance, indicating that the normal distribution is not a good fit to the data. Table 2 contains the results of fitting a lognormal distribution. Again, except for the chi-square test, all other tests have the same conclusion, which is to not reject the lognormal distribution. It should be noted that because of the way the intervals were defined, the

```
data a;
do j = 1 to 100;
    y=3+sqrt (.5)*rannor (934410);
    x=expr(y);
    output;
end;
run;

proc means;
var x y ;
run;

/* goodness of fit test of X observations*/
proc capability data=a;
var x ;
histogram /
midpoints=0 30 60 90 120 150 180
lognormal (l=1 color=red)
normal      (l=8 color=yellow)
cframe = ligr;

run;

/* goodness of fit test of Y observations*/
proc capability data=a;
var y;
histogram /
midpoints=1 2 3 4 5 6
lognormal (l=1 color=red)
normal      (l=8 color=yellow)
cframe = ligr;

run;
```

Figure 1 SAS Program

chi-square test will not be valid because intervals for larger values of X are mostly empty and have expected values less than 5. In fact, SAS sends out

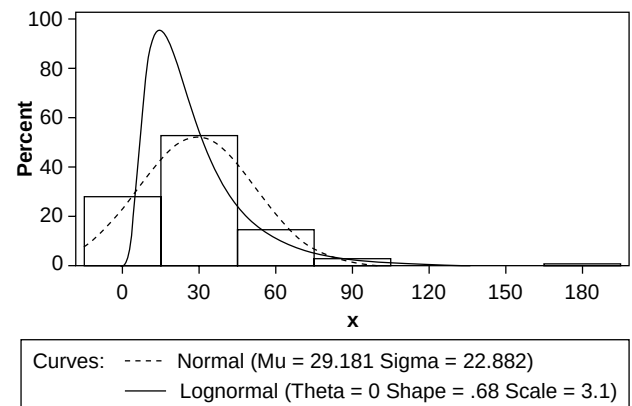


Figure 2 Histogram of Data With Fitted Normal and Lognormal Densities

Table 1 SAS Output: Goodness-of-Fit Tests on X

Test	Statistic		DF	p Value	
Kolmogorov-Smirnov	D	0.14540565		Pr > D	<0.010
Cramer-von Mises	W-Sq	0.74062491		Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.51011616		Pr > A-Sq	<0.005
Chi-Square	Chi-Sq	3.66244120	4	Pr > Chi-Sq	0.454

Note: H_0 : normal distribution is a good fit.

Table 2 SAS Output: Goodness-of-Fit Tests on X

Test	Statistic		DF	p Value	
Kolmogorov-Smirnov	D	0.05567901		Pr > D	> 0.150
Cramer-von Mises	W-Sq	0.05199648		Pr > W-Sq	0.486
Anderson-Darling	A-Sq	0.34639466		Pr > A-Sq	0.484
Chi-Square	Chi-Sq	9.54984429	4	Pr > Chi-Sq	0.049

Note: H_0 : lognormal distribution is a good fit.

Table 3 SAS Output: Goodness-of-Fit Tests on Y

Test	Statistic		DF	p Value	
Kolmogorov-Smirnov	D	0.05567901		Pr > D	> 0.150
Cramer-von Mises	W-Sq	0.05199648		Pr > W-Sq	> 0.250
Anderson-Darling	A-Sq	0.34639466		Pr > A-Sq	> 0.250
Chi-Square	Chi-Sq	3.71521303	2	Pr > Chi-Sq	0.156

Note: H_0 : normal distribution is a good fit.

empty nor have expected values less than 5.

Table 3 displays the results of testing the normality assumption on Y . As expected, the normal fit was not rejected. Even the chi-square test, which is still not valid in this case because of some intervals with expected values less than 5, has the same conclusion as the other tests. Note that Y may be obtained from X via the transformation $Y = \log X$. This technique of transforming data is commonly used in applications as a remedy to get normal data from nonnormal data.

—Inmaculada Aban
and Edsel Pena

See also Chi-Square Test for Goodness of Fit; Chi-Square Test for Independence

Further Reading

D'Agostino, R. B., & Stephens, M. A. (Eds.). (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker.

Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., & Mesbah, M. (Eds.). (2002). *Goodness of fit tests and model validity*. Papers from the International Conference held in Paris, May 29–31, 2000. Boston: Birkhäuser Boston.

NIST/SEMATECH e-Handbook of Statistical Methods: <http://www.itl.nist.gov/div898/handbook/>

the following message as a reminder: “The chi-square statistic has been computed using expected values less than 1. You can regroup the data using the MIDPOINTS= option.” Because SAS does not allow unequal interval size, one has to use other programs to perform the chi-square test using carefully defined intervals so that there will be no intervals that are

GRADUATE RECORD EXAMINATIONS

The Graduate Record Examinations (GRE) test, published by Educational Testing Service, is a

standardized test to be taken prior to applying for graduate education. The College Board was originally designed to develop and implement standardized tests for the admission and placement of undergraduate students. After the success of the Scholastic Aptitude Test (SAT) in determining a student's readiness for postsecondary education, the College Board went a step higher, developing a series of tests for graduate admission in 1936, which included the GRE.

Presently, the GRE is designed to supplement undergraduate achievements, including grade point average (GPA), providing graduate admission committees a common measure from which to compare the qualifications of applicants. Many graduate programs have an explicit cut-off for applicant GRE scores and rarely admit students scoring below these levels.

The GRE offers two different tests for students, a General Test and a Subject Test, both of which may be required for applying to an accredited graduate or professional program. The GRE General Test produces a different score for each of its three sections:

1. The analytic writing section measures the student's ability to articulate and evaluate meaningful, supported arguments. This section focuses more on critical thinking than on the basic mechanics of language.
2. The verbal reasoning section measures the student's ability to analyze written material. Test takers are expected to synthesize information and recognize relationships between words, sentences, and overarching concepts.
3. The quantitative reasoning section measures the student's ability to solve problems using the basic concepts of arithmetic, algebra, geometry, and data analysis.

The GRE Subject Test evaluates undergraduate achievement in eight specific areas of study: Biochemistry/Cell and Molecular Biology, Biology, Chemistry, Computer Science, Literature in English, Mathematics, Physics, and Psychology.

Sternberg and Williams evaluated the GRE's ability to predict graduate marks for psychology students at Yale University. The GRE General Test predicted

graduate grades only modestly (correlation of .17), whereas the GRE Subject Test in psychology predicted marks more strongly (correlation of .37). The Graduate Record Examinations Board has reported that when combined, undergraduate GPA, the GRE General Test, and the GRE Subject Test strongly predicted (correlation of .50) first-year grades in graduate school for psychology students. Beyond predicting first-year graduate marks, however, GRE scores are not found to be useful for predicting other aspects of graduate performance, including the ratings of analytical, creative, research, and teaching abilities by primary advisers, and the ratings of dissertation quality by faculty members. Therefore, it appears that the GRE should be taken into consideration along with other indicators of qualification, including undergraduate GPA and past experience, when evaluating applicants to graduate studies.

—John R. Reddon and Michelle D. Chudleigh

See also Educational Testing Service

Further Reading

Graduate Record Examinations Board. (1997). *GRE 1997–98: Guide to the use of scores*. Princeton, NJ: Educational Testing Service.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? *American Psychologist*, 52, 630–641.

Graduate Record Examinations Web site: <http://www.gre.org>

GRAND MEAN

The grand mean is most often used in the computation of the F value in analysis of variance, or ANOVA. It is the overall mean from which group means are subtracted when computing the between-group variance. The between-group variance and the within-group variance are the two variance estimates that are used to create the F ratio. The grand mean can also be used as an estimate of the average of all the scores in a group.

Table 1 Sample Data for Computing the Means and Grand Mean

	Group 1	Group 2	Group 3
	6	7	8
	4	6	7
	5	5	6
	6	3	5
	5	4	6
	4	5	5
	3	3	4
	4	2	4
	5	3	3
	6	3	3
Group Average	4.8	4.1	5.1
Grand Mean	4.7		

For example, Table 1 shows three sets of 10 scores along with the mean for each of the groups. The grand, or overall, mean is shown in this table as well.

—Neil J. Salkind

See also Analysis of Variance (ANOVA); Average; Mean

GRAPHICAL STATISTICAL METHODS

The use of diagrams as a means for summarizing and analyzing data has a long tradition in data exploration. Long before the discipline of statistics was established, people saw the need to summarize available data. Maps are probably the earliest diagrams used to depict and pass on information. The 19th century saw a surge in the creation of beautiful artistic charts with a more data-oriented background. Some of the most influential persons creating and establishing statistical graphics were Charles Joseph Minard, Dr. John Snow, Florence Nightingale, and William Playfair.

Charles Joseph Minard (1781–1870) was a French engineer. He is

best known for his rich portfolio of intricate maps. His most famous piece of work is the chart of Napoleon’s March to Moscow (see Figure 1), which, according to Tufte, is the best statistical graphic ever drawn. Minard manages to artistically incorporate complex data in a single chart: size of the army, marching direction, spatial location, temperature, and the dates of river crossings tell the sad story of the deaths of hundreds of thousands of French soldiers.

Dr. John Snow (1813–1858) is one of the founding fathers of modern epidemiology. During the 1854 cholera outbreak in central London, Dr. Snow used a map (see Figure 2) to mark all outbreaks and put them in geographical reference to the water pumps regularly used by the victims. In this way, he was able to identify the pump near Broad Street as the source of the epidemic.

William Playfair (1759–1823) was a Scottish inventor and writer. He introduced some of the chart types still in use today, such as the pie chart, the bar chart, and line diagrams to depict time series. In 1786, he published *The Commercial and Political Atlas*, which contained 43 time series plots and one bar chart. One of the most famous charts is shown in Figure 3, describing the relationship of food prices and wages throughout the reigns of Elizabeth I to George IV.

Florence Nightingale’s (1820–1910) most famous diagram discusses insufficient sanitary conditions in military field hospitals during the Crimean War,

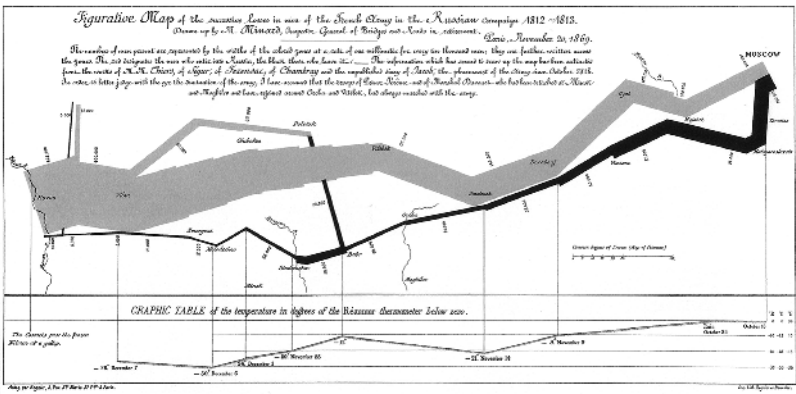


Figure 1 Minard’s Chart of Napoleon’s 1812 March on Moscow



Figure 2 Snow's Map of Central London Investigating the Cause for the Cholera Epidemic of 1854

which led to the (preventable) deaths of thousands of soldiers. She invented polar-area diagrams, where the statistic being represented is proportional to the area of a wedge in a circular diagram. In 1858, Florence Nightingale was elected the first female member of the Royal Statistical Society, and she became an honorary member of the American Statistical Association

introduced a variety of now well-known and used diagram types, such as the box plot (box-and-whisker plot) and the (slightly outdated) stem-and-leaf plot (see Figure 5 for examples). The ground-breaking new features of Tukey's inventions are their abstractness. Data points are plotted along axes that do not have a direct relation to space or time. In this respect, Tukey is the founding father of statistical graphics as it is seen today. We use visualization methods to portray abstract relationships among variables. This is the essential difference between statistical graphics and other areas of information visualization, which are principally concerned with the rendering of objects and phenomena in physical 3-D space.

Today, the standard ensemble in the toolkit of a statistician consists of the following:

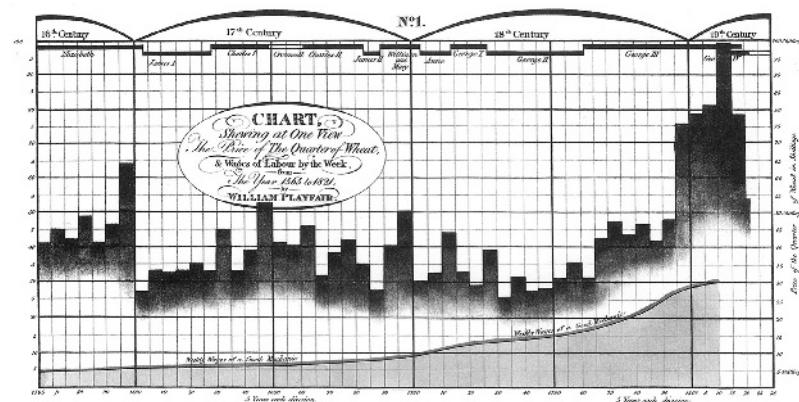


Figure 3 Playfair's "Prices, Wages, and Reigns"

- one-dimensional plots: bar charts and pie charts for categorical variables, histograms, dot plots, and box plots for continuous variables
- two-dimensional plots, such as the scatter plot for continuous variables, and combination of one-dimensional plots for a mixture of one continuous variable and categorical variables
- mosaic plots for displaying multivariate categorical data
- rotating plots for three dimensions
- projection techniques for higher dimensional data
- interactive tools, such as selection and linked highlighting to gain insight in higher dimensional structures in the data

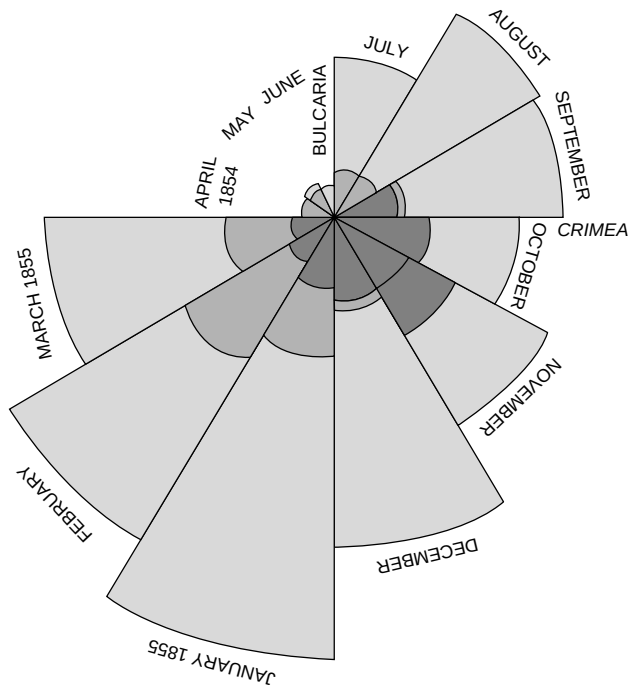


Figure 4 Nightingale's Polar-Area Diagram of the Causes of Mortality in the Army in the East

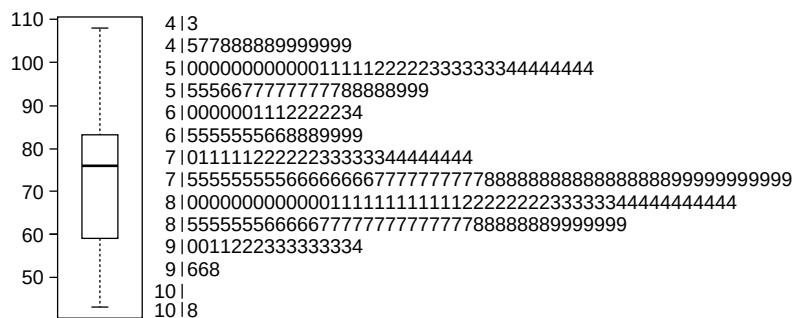


Figure 5 Box Plot (left) and Stem-and-Leaf Plot (right) of the Time Between Eruptions of Old Faithful

This list is certainly not complete—there are variations of each of these diagrams under various names; there are other types of diagrams; and there are new kinds of data, such as data streams or (Internet) network data, which cannot be displayed well by any of the diagrams mentioned above. Some of the ongoing research in graphics deals with the problem of how best to display new types of data. Other research deals with how best to present information and how to come up with “good” (i.e., faithful) graphics, because bad graphics seem to appear everywhere, particularly in nonstatistical environments. Ground-breaking work has been published by Edward R. Tufte in this area. In *The Visual Display of Quantitative Information*, Tufte describes a set of rules for preserving and checking graphical integrity in charts.

The *Semiology of Graphics* (first edition in French in 1967) by Jean-Jacques Bertin provides a summary of elements of graphics and conceptual principles. Leland Wilkinson's monograph *The Grammar of Graphics* can be seen as a logical successor to Bertin, even though the book is self-reliant. Wilkinson describes elements of graphics and presents an algebra for them that allows a flexible and consistent way of constructing and describing graphs.

Interactive Statistical Graphics

With the rise of computer technology and general availability of personal computers on almost everybody's desk, software for statistical graphics became widely accessible. This led statisticians to search for ways of "interacting" with their data. The collection

of movies in the ASA Statistical Graphics Section Video Lending Library paint an impressive picture of the developments in statistical graphics since 1960. But what makes software interactive? There is an almost bewildering abundance of applications that go under the heading of interactive software, yet there seem to be quite different opinions of what interactivity means. The definition and use of this term is not quite clear, even

among computational statisticians, as a survey on this topic by Swayne and Klinke made clear.

Here, we will refer to human-computer interaction based on the definition proposed by Unwin as the direct manipulation of plots and plotting elements in them. This goes back to one of the first definitions of interactivity made by Becker and Cleveland and Becker et al.: “dynamic methods have two important properties: direct manipulation and instantaneous change.” The data analyst takes an action of an input device and something happens, virtually instantaneously, on a computer graphics screen. Huber (1988) corrects the term of dynamic graphics to high-interaction graphics. “Virtually instantaneously” is often interpreted as real time changes, with the maximum response time set to 20 ms or, equivalently, an update speed of 20 frames per second. This, however, looks rather dangerous because it emphasizes the role of the underlying hardware and may even lead to different decisions about which methods can be classified as interactive. Therefore, it is probably better to speak of a potentially interactive method if it fulfills the proposition of being directly manipulative. Direct manipulation depends on two conditions:

1. Immediacy of place—by using a pointing device such as a mouse, the analyst can specify visually the areas of the plot, which are meant as a starting point of an action.
2. Immediacy of action—the action is triggered by using a clicking device such as a mouse, pressing keys on the keyboard (but not typing in commands), or via some other input device.

Interactive methods let graphics become real tools of data analysis. The most commonly used methods are linked highlighting, brushing, identifying, and zooming. Brushing was first introduced by Becker et al. as a tool for identifying and cross-linking points in scatterplot matrices. The idea of brushing is to mark all points inside the brushing area, usually a rectangle, and mark corresponding values in other graphs in the same way. Moving the brushing area to different positions leads to changes of marked points in all other graphs, revealing relationships between variables. The most commonly used brushing technique is highlighting. Figure 6 shows an example of linked highlighting in the iris data set: The brush is

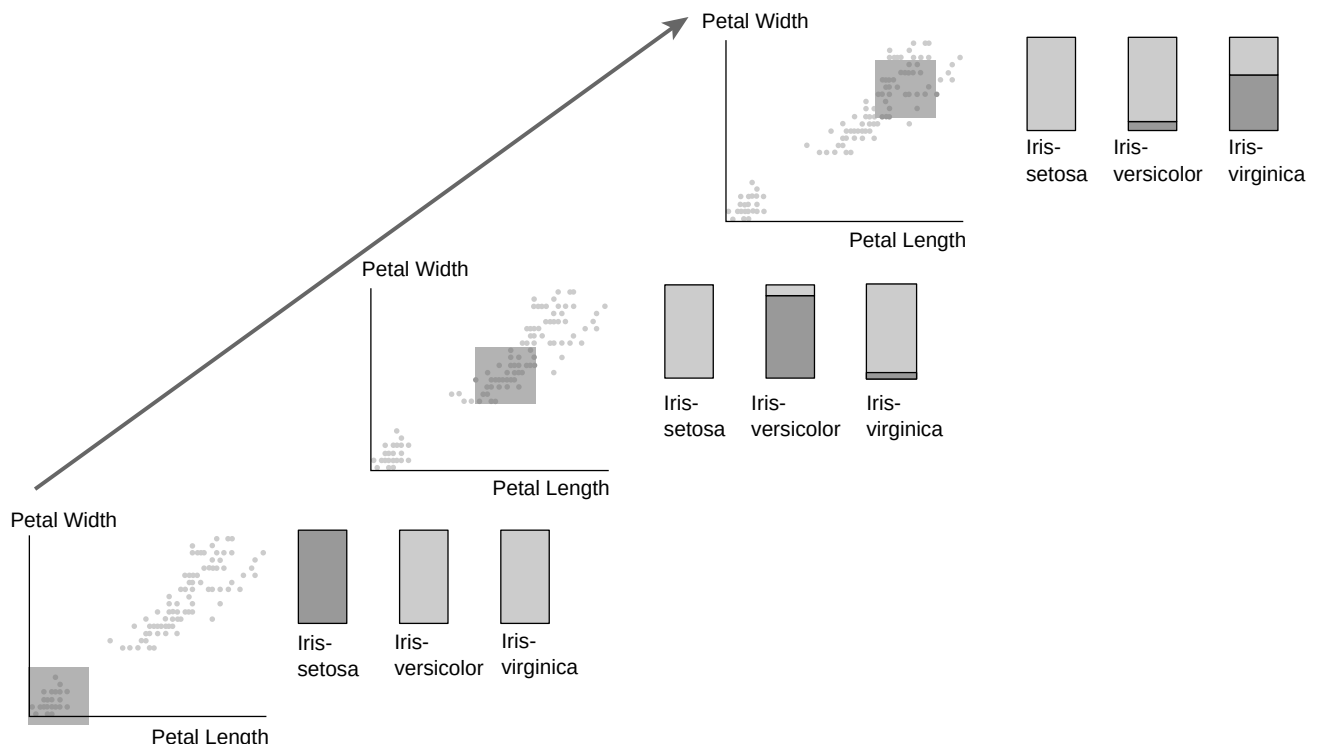


Figure 6 Example of Linked Highlighting in the Iris Data Set

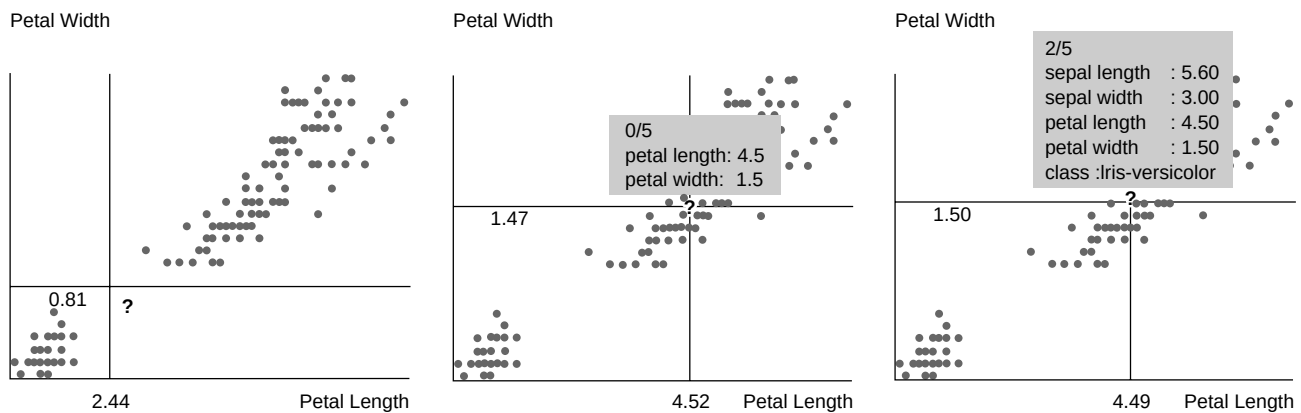


Figure 7 Examples of Identifying Objects in Scatterplots

moved along the diagonal of a scatterplot showing petal length versus petal width. Corresponding values are highlighted by color in a linked bar chart of iris type. As the brush moves from bottom left to top right, iris species are *setosa*, then predominantly *versicolor*, and finally *virginica*.

Being able to identify objects and individuals in graphs is crucial. Interactive querying allows the output of context-sensitive information. This operation is normally triggered by a movement of the mouse or by point and click of a graphical object such as an axis, a point, or a bar. Figure 7 shows examples of identifying objects in scatterplots. Query-clicking in empty space (left) gives the coordinates along the x- and y-axes. Query-clicking points triggers output of the number of observation (five, in the example) and their coordinates. More in-depth information is shown on the right: All available information for Observation #2 pops up.

Of course, all interactive methods depend highly on the specific implementation of a particular software. On the up side, though, basically all interactive systems now do have implementations of these crucial interactive methods even though names and functionality may vary slightly.

The final interactive, which seems to be common to all interactive systems for data exploration, is logical zooming. Whereas standard zooming enlarges the displayed graphical elements, logical zooming works on the underlying model and changes it to display

more details. Logical zooming is quite natural when working with maps. Starting with a country map, we zoom into a regional map, a city map, and finally a street map, which shows us the neighborhood in every detail. This gives us a tool for breaking down large data sets into smaller parts, which are easier to analyze.

—Heike Hofmann

See also Area Chart; Bar Chart; Line Chart; Pie Chart

Further Reading

- Becker, R. A., & Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29, 127–142.
- Becker, R. A., Cleveland, W. S., & Weil, G. (1988). The use of brushing and rotation for data analysis. In W. S. Cleveland & M. E. McGill (Eds.), *Dynamic graphics for statistics* (pp. 247–275). Pacific Grove, CA: Wadsworth.
- Friendly, M., & Denis, D. J. (2006). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Retrieved from <http://www.math.yorku.ca/SCS/Gallery/milestone/>
- Nightingale, F. (1858). *Notes on matters affecting the health, efficiency, and hospital administration of the British Army, founded chiefly on the experience of the late war*. London: Harrison and Sons.
- Playfair, W. (1786). *The commerical and political atlas*. London: Debrett.
- Swayne, D., & Klink, S. (1999). Introduction to the special issue on interactive graphical data analysis: What is interaction? *Journal of Computational Statistics*, 1, 1–6.

- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Unwin, A. R. (1999). Requirements for interactive graphics software for exploratory data analysis. *Journal of Computational Statistics*, 1, 7–22.

Statistical graphics video lending library: <http://www.amstat-online.org/sections/graphics/library.php>

GRESHAM, FRANK M. (1949–)

Frank Gresham was born in the small town of Greenville, South Carolina, in April 1949. As a young man growing up in Simpsonville, South Carolina, Frank was a good football player and was offered a scholarship to play football for a small local college. However, Frank had other plans and left home to pursue his college education.

He earned his Bachelor of Science degree in psychology from Georgia State University in 1973. He then went on to the University of South Carolina (USC), where he earned his M.Ed. in Rehabilitation Counseling. While working as a counseling coordinator for the South Carolina Department of Corrections, he went on to earn his PhD in psychology from USC.

After receiving his PhD, Dr. Gresham accepted a position at Iowa State University. Two years later, he was asked to become the director of the School Psychology Program at Louisiana State University. Dr. Gresham held this position for 8 years and published dozens of articles in the areas of applied behavior analysis, social skills, and behavioral consultation. As a result of his extraordinary research, he was given the Lightner Witmer Award in 1982. The highly coveted Lightner Witmer Award is given by the American Psychological Association (APA) for outstanding research contributions by a school psychologist. In 1985, he was recognized as a Fellow by both the APA and the APA's Division of School Psychology.

In 1989, Dr. Gresham accepted a position as the director of the Combined Clinical and School Psychology Program at Hofstra University. His most notable accomplishment during his 2 years at Hofstra University was to co-author the Social Skills Rating System (SSRS). The SSRS is used frequently by school psychologists throughout the United States and abroad for the assessment of children experiencing social, emotional, and behavioral difficulties. In 1991, Dr. Gresham accepted a position as the director of the School Psychology Program at the University of California, Riverside (UCR). During his 14 years at UCR, Dr. Gresham continued to publish research articles and chapters at an unparalleled rate. He was the recipient of eight federally funded grants for the study of learning disabilities, literacy, and emotional and behavioral disorders, and he received many honors for his scholarly work and research, including the Senior Scientist Award by the APA and the rank of Distinguished Professor.

Currently, the field of school psychology is undergoing a major paradigm shift in the identification of children with learning disabilities. The Response-to-Intervention model proposed by Dr. Gresham to the U.S. Office of Special Education Programs in 2001 was recently approved by the U.S. Congress to replace the traditional discrepancy model used for identifying learning disabilities for the past 30 years.

In addition to his work as a researcher, teacher, and mentor, Dr. Gresham was appointed as an expert witness to the President's Commission on Excellence in Special Education. Dr. Gresham is also a consultant to state psychological associations in more than 45 states, as well as Canada and Australia.

Dr. Frank Gresham continues to be one of the most respected and prolific scholars of our time in the field of school psychology.

—Alberto Restori

Further Reading

Frank Gresham home page: <http://www.behavioralinstitute.org/Frank%20Gresham.htm>

GROUNDING THEORY

Grounded theory is a broad perspective on how to conduct qualitative social science research. It comprises a distinctive methodology, a particular view of scientific method, and a set of procedures for analyzing data and constructing theories. The methodology provides a justification for undertaking qualitative research as a legitimate, indeed rigorous, form of inquiry. The conception of scientific method depicts research as a process of inductively generating theories from closely analyzed data. The specific procedures used in grounded theory comprise an array of coding and sampling procedures for data analysis, and a set of interpretative procedures that assist in the construction of theory that emerges from, and is grounded in, the data. In all of this, grounded theory researchers are expected to meet the canons of doing good scientific research, such as reproducibility and generalizability.

Grounded theory has been employed by researchers in a variety of disciplines, including sociology, nursing studies, education, management science, and psychology. It is probably the best known and widely used qualitative research methodology available today.

History

The grounded theory method was introduced in the 1960s by two American sociologists, Barney Glaser and Anselm Strauss, and has been further developed by them and others. Grounded theory was introduced to serve three purposes. First, it endeavored to close the gap between theory and empirical research by having theory emerge from the data. Second, it began to spell out the inductive logic involved in producing grounded theory. Finally, it provided a justification for the careful and rigorous use of qualitative research methods in sociology.

Deriving its theoretical underpinnings from the philosophy of American pragmatism and the related social theory, symbolic interactionism, grounded theory portrays research as a problem-solving endeavor concerned with understanding action from the perspective of the human agent. Strauss was

heavily influenced by the University of Chicago tradition in qualitative social research, with its emphasis on the method of comparative analysis and the use of participant observation. Glaser was strongly influenced by the quantitative research tradition at Columbia University, and he brought to grounded theory important ideas from this tradition and translated them into qualitative terms.

Both Glaser and Strauss continued to develop the methodology of grounded theory, although in separate publications. From the 1980s onwards, their formulations of grounded theory diverged somewhat. Glaser sees himself as having remained true to the original conception of grounded theory, with its emphasis on studying basic social processes, the use of the constant comparison method, and the formulation of theories by letting abstract relationships between theoretical categories emerge from the data. Strauss, in association with Juliet Corbin, developed new methods of analysis in place of the strategy of constant comparative analysis, and they stressed the importance of verification of theory as well as its generation. Glaser has strongly objected that Strauss and Corbin's approach forces data and their analysis into preconceived categories instead of letting the categories emerge from the data. Although Strauss acknowledges that there are differences, he maintains that both he and Glaser advocate use of the same basic procedures for doing grounded theory research.

Philosophical Perspectives

Grounded theory has also been presented from a number of different philosophical positions. Glaser adopts a general empiricist outlook on inquiry. This has sometimes been described by commentators as "positivism." However, given the influence of pragmatism on his early formulations of grounded theory, this is an unfair characterization. Strauss's own characterization of grounded theory leans toward a social constructionist perspective. Kathy Charmaz has provided an explicitly constructivist depiction of grounded theory that breaks with the objectivism of Glaserian grounded theory. On a constructionist perspective, social reality is not revealed so much as socially constructed in the course of inquiry. David

Rennie offers a hermeneutic interpretation of the grounded theory method that is able to provide an understanding of the meaning of text and reconcile the tensions that exist between realism and relativism in orthodox accounts of the method. Finally, Brian Haig offers a reconstruction of grounded theory methodology from a broadly scientific realist perspective. On this interpretation, grounded theory method involves the inductive discovery of empirical phenomena followed by the abductive construction of theory to explain the phenomena.

Procedures

The variety of interpretations of grounded theory extend to characterizations of the method itself. In efforts to identify empirical social phenomena, and construct theories that explain those phenomena, almost all accounts of grounded theory adopt the three major research strategies of data coding, memo writing, and theoretical sampling. In grounded theory, data gathering and data analysis are interactive: From the time data collection begins, grounded theorists engage in data analysis, which leads to further data collection, subsequent data analysis, and so on.

The first data-analytic phase of grounded theory begins with the coding of data. This is undertaken to conceptualize the data by discovering categories into which they fit. The coding process has three phases: open coding, axial coding, and selective coding. In open coding, researchers describe the data by looking at them line by line. This strategy of focusing on small units of data, and their interpretation, encourages the development of a theoretical sensitivity to new ideas about the data and helps prevent the forcing of data into existing categories. Strauss and Corbin maintain that when a full array of categories has been identified, one should undertake axial coding, whereby one puts the data back together again in new ways by making connections between the numerous categories. After that, a selective coding step is implemented in which the researcher looks to systematically identify those categories that relate closely to the core category. The core category lies at the heart of the emerging theory and is central to the theory's integration.

Although memo writing can occur at any stage of the research process, it frequently takes place between the coding of data and the writing of the initial draft of the research report. Memos are written to identify, develop, and keep track of theoretical ideas. Where relevant, they are recorded, recalled, and reworked to produce new theoretical memos. Memo writing becomes more systematic, focused, and intense as theory of greater density and coherence is produced.

Memos written about data codes and theoretical ideas enable the researcher to identify gaps that require the collection of further data. For this, theoretical sampling is undertaken. With theoretical sampling, in contrast with traditional representative sampling, decisions about which data to collect, code, analyze, and interpret are directed by the emerging grounded theory. Theoretically relevant events, activities, and populations are all sampled, and the comparisons between these are aimed at increasing the conceptual density and integration of the emerging theory. Thinking effectively about data in theoretical terms requires an adequate degree of theoretical sensitivity. When the additional gathering and analysis of data no longer contribute to the understanding of a concept or category, a point of theoretical saturation is reached. At this point, one stops collecting data in respect of a category and moves to consider another category or concept.

Grounded theory considers writing to be an important part of the research process. This extends beyond the writing of memos to writing up the research report itself. One of the major goals in drafting the research report is to present a fully integrated account of the phenomena studied. This will involve highlighting areas that are insufficiently integrated and working to remedy these through multiple drafts if needs be. Grounded theory provides a number of rules of thumb, or heuristics, to improve the integrative value of the research report.

Criticisms

Despite its popularity, grounded theory has been subjected to a number of criticisms. One criticism asserts that grounded theory is a regression to a simple "Baconian" form of inductive science. In this

interpretation, grounded theory is depicted as a tabula rasa view of inquiry, which maintains that data analysis and interpretation are not dependent on concepts or theories. However, this is an unwarranted criticism. In their first book on grounded theory, Glaser and Strauss explicitly stated that the researcher must have a perspective in order to discern relevant data and abstract relevant categories from them. In their view, the researcher seeks to obtain emergent diverse categories at different levels of abstraction by bracketing potentially relevant existing facts and theories for some time.

A further criticism of grounded theory is the claim that the reasoning involved in the generation of grounded theory is not inductive, as Glaser and Strauss claim, but abductive. Inductive reasoning is typically a generalizing inference, and it is difficult to see how such descriptive inferences could lead to the causes that explain generalizations. In contrast, abductive inference is explanatory inference, often from presumed effects to underlying causes. It is this type of reasoning process that leads from facts to explanatory theories. It is surprising that the originators of grounded theory have not appealed to abductive reasoning, given its prominence in the work of the pragmatist tradition from which they have drawn.

Yet another criticism of grounded theory points out that its methodology stresses the importance of theory generation at the expense of theory verification, or validation. However, whereas the first writings on grounded theory method deemphasized theory validation in favor of theory generation, this was in part due to Glaser and Strauss's desire to break from the hypothetico-deductive emphasis on theory testing that dominated 20th-century sociology. Glaser has continued to see grounded theory primarily as a theory generation method, but Strauss has come to emphasize the importance of theory verification in grounded theory research.

Although grounded theory does not articulate a precise account of the nature of theory testing, some writings on the method make it clear that there is more to theory appraisal than testing for empirical adequacy. Clarity, consistency, parsimony, density, scope, integration, fit to data, explanatory power, predictiveness, heuristic worth, and application are all

mentioned by Glaser and Strauss as relevant evaluative criteria, although they do not elaborate on these, nor do they work them into an integrated view of theory appraisal.

Conclusion

Grounded theory methodology continues to be the subject of critical epistemological examination. Its methods continue to be employed widely, both in full and in part, in social science research, especially with the aid of computer programs for qualitative data analysis. Although initially developed as an approach to qualitative research, the use of grounded theory method in the future is likely to employ a mix of qualitative and quantitative research methods and to link with other methods that give explicit emphasis to the construction of theory that is undertaken to explain the data patterns obtained about empirical social phenomena.

—Brian Haig

See also Authenticity

Further Reading

- Charmaz, K. (2000). Grounded theory: Constructivist and objectivist methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509–535). Thousand Oaks, CA: Sage.
- Dey, I. (1999). *Grounding grounded theory*. San Diego, CA: Academic Press.
- Glaser, B. G. (1992). *Emerging versus forcing: Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Haig, B. D. (1995). *Grounded theory as scientific method* (Philosophy of Education Society Yearbook 1995, pp. 281–290). Urbana: University of Illinois Press.
- Rennie, D. L. (2000). Grounded theory methodology as methodological hermeneutics: Reconciling realism and relativism. *Theory and Psychology*, 10, 481–502.
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.
- Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Thousand Oaks, CA: Sage.

Grounded theory resources: http://dmoz.org/Science/Social_Sciences/Methodology/Grounded_Theory/

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Rennie, D. L. (2000). Grounded theory methodology as methodical hermeneutics: Reconciling realism and relativism. *Theory and Psychology*, 10(4), 481–502.

David Rennie argues that the realism-relativism duality addressed by the grounded theory approach to qualitative research is best accounted for when the method is understood to be an inductive approach to hermeneutics. Phenomenology, C. S. Peirce's theory of inference, philosophical hermeneutics, pragmatism, and the new rhetoric are drawn upon in support of this argument. It is also held that this formulation of the **grounded theory** method opens the possibility that the method improves on earlier approaches to methodical hermeneutics. As an outcome of this formulation, the debate on the validity and reliability of returns from the grounded theory approach is cast in a new light. The new methodical hermeneutics is discussed in terms of prior attempts to relate hermeneutics to method.

GUTTMAN SCALING

Guttman scaling (also called *scalogram analysis*) is a method of scaling involving items that reflect increasing levels of extremity on a single dimension of interest. This procedure was developed by Louis Guttman and was introduced in response to the concern that prior attitude measures might sometimes tap multiple constructs. For instance, a scale might include (among others) the statements presented in Table 1.

These statements both represent negative views toward obesity. Item 1 is the most extreme, and it seems reasonable to expect that individuals who endorse it might also endorse Item 2. However, it is

Table 1 Obesity Items Reflecting Two Dimensions

Items

Obesity is a serious health threat that leads to early death.
Obese individuals are less physically attractive than people who aren't overweight.

plausible that an individual might have an attitude toward obesity based solely on the associated health risks. This person might agree with the more extreme (health-related) item, but disagree with the less extreme attractiveness item. A scale containing these types of items would be assessing more than one construct. Therefore, it would be difficult to determine whether two people receiving the same score actually reported the same attitude, or whether their scores were driven by different dimensions. Guttman's approach was designed to correct this potential problem.

Guttman Scale Criteria

A Guttman scale is composed of a set of statements to which respondents indicate their level of agreement. Because these scales are meant to assess a single dimension, items are selected to vary on extremity alone. Therefore, statements chosen must be unipolar, reflecting gradations of either support or rejection. It is expected that individuals will agree with all items leading up to their most extreme endorsement and will disagree with items of higher extremity. With this goal in mind, researchers aim to select items that will fit a stepwise sequence of extremity. In fact, successful Guttman scales often consist of nested statements, such that agreeing with any one item almost necessitates agreement with items of lower extremity. This is a useful strategy because individuals with extreme attitudes might otherwise reject more moderate items on the basis that they denote too weak a stance. For example, the statements presented in Table 2 might be used to examine respondents' attitudes toward the film *Casablanca*.

Table 2 Sample Guttman Scaling Items

Items

Casablanca is the greatest movie ever made. (most extreme)
Casablanca is one of the top 5 movies of all time.
Casablanca is an exceptional film, one of the best.
Casablanca is a great film, well above average.
Casablanca is a good movie. (least extreme)

To satisfy the requirements of a Guttman scale, individuals who agree with Item 1 should also agree with all other statements listed because they represent less extreme views. Likewise, respondents who disagree with Item 1 but agree with Item 2 should endorse all remaining items, and so on. This not only ensures that scores obtained are interpretable (because individuals with the same score will have endorsed the same statements), but it also introduces the possibility that knowledge of individuals' scores alone will allow exact replication of their responses. This "reproducibility" is a key element of Guttman scaling.

Evaluating Reproducibility

To assess the criterion of reproducibility, researchers first organize their data by creating a matrix in which columns represent scale items and rows represent respondents. If an individual agrees with a statement, a 1 is placed in the cell common to both the item and the respondent, and if a person disagrees with a statement, a 0 is placed in that cell. Individuals' scores are calculated by counting the number of 1s in each row. If this type of matrix were created for the *Casablanca* items mentioned earlier, it might look like Table 3. This table indicates, for example, that Persons 3 and 9 agreed with all items, whereas Person 6 agreed only with Item 1.

Table 3 Initial Response Matrix (Items by Respondents)

Person #1	Items					Total Score
	2	3	4	5	6	
1	0	0	0	1	1	2
2	0	0	1	1	1	3
3	1	1	1	1	1	5
4	1	0	0	1	1	3
5	0	0	0	0	1	1
6	1	0	0	0	0	1
7	0	0	0	1	1	2
8	0	1	1	1	1	4
9	1	1	1	1	1	5
10	0	1	1	1	1	4

After creating this preliminary matrix, rows and columns must then be rearranged according to the number of 1s they contain. This is done such that the column with the most 1s is placed at the far right and the row with the most 1s is at the top (see Table 4). This matrix is used to identify items of equivalent extremity and to determine which items produce inconsistent response patterns.

In a perfect Guttman scale, respondents would agree with all statements leading up to their most extreme endorsement, overall response patterns would support the hypothesized extremity of each statement, and all individuals with a given score would agree and disagree with the same statements. Furthermore, because scores would be entirely dependent on the extremity of the statements endorsed, the data from such a scale would be completely reproducible from the scores alone. In practice, such a pattern is seldom achieved. For example, the data presented in Table 4 are not consistent with a perfect Guttman scale. In fact, both Person 4 and Person 6 endorse the most extreme statement (Item 1) and subsequently disagree with items deemed lower in extremity. Individuals who break the pattern by responding unexpectedly are said to have committed errors. Violations of this type serve to undermine reproducibility because response patterns can no

Table 4 Reordered Response Matrix, Including Errors

Person #	Items					Total Score	Errors
	1	2	3	4	5		
3	1	1	1	1	1	5	0
9	1	1	1	1	1	5	0
8	0	1	1	1	1	4	0
10	0	1	1	1	1	4	0
2	0	0	1	1	1	3	0
4	1	0	0	1	1	3	1
1	0	0	0	1	1	2	0
7	0	0	0	1	1	2	0
5	0	0	0	0	1	1	0
6	1	0	0	0	0	1	1

longer be determined from scores alone. Because reproducibility is a proxy for unidimensionality in Guttman scaling, procedures have been devised to assess the degree to which a set of responses violates the ideal pattern.

The coefficient of reproducibility (defined as $[1 - \text{total number of errors} / \text{total number of responses}]$) was suggested by Guttman in 1950 and is used to indicate the level of reproducibility. Guttman proposed a coefficient of .90 as the level needed to assume unidimensionality. The total number of responses for a scale is equal to the number of respondents multiplied by the number of items, but the number of errors is ambiguous initially, so determining the total error count is more complicated. A limited number of response sets fit the pattern required for Guttman scaling, and it is necessary to establish which of these a person “should” have given to determine a person’s number of errors. Errors are thus conceptualized as the number of responses that need to be changed to create one of the acceptable patterns. This “intended” response pattern is selected to optimize reproducibility, so each individual is charged with the fewest errors possible.

For example, in Table 4, one break from the ideal pattern occurs with Person 6. This individual agreed with the most extreme item, but disagreed with all statements deemed lower in extremity. This response set can be explained in several ways. It could be assumed that because the individual endorsed the most extreme item, he or she should also have agreed with the remaining items (resulting in four errors). However, the pattern can also be explained with fewer errors if the endorsement of Item 1 is considered a mistake. This assumption charges only one error to Person 6 and so is the appropriate choice.

Researchers have questioned the merit of Guttman’s reproducibility coefficient because large coefficients are not necessarily evidence for unidimensionality. For instance, the reproducibility of a statement with two response options (e.g., agree/disagree) cannot be less than the proportion of people who gave the most popular response for that item. That is, if 95 out of 100 respondents agree with a statement, the maximum number of errors for that item is 5, so the smallest possible reproducibility for

the statement would be .95. Therefore, if several items are selected to which a large percentage of respondents agree (or disagree), it is conceivable that a scale could be highly reproducible for reasons unrelated to content. To remedy this, a more precise alternative, the error ratio, has been proposed to assess reproducibility. This ratio is calculated by dividing the number of observed errors by the maximum number of errors possible, so that the ratio will range from 0.00 (no observed errors) to 1.00 (maximum error observed). This index is more sensitive to the absence of unidimensionality, with higher ratios indicating lower reproducibility.

Strengths and Weaknesses

The major asset of Guttman scaling is its focus on unidimensionality. Relative to other methods, Guttman scales are more likely to succeed in tapping only a single construct. Hence, scores often have more straightforward interpretations. However, several drawbacks are also evident. First, because increasing the number of items makes it more difficult to satisfy Guttman criteria, scales are necessarily short. Hence, variability among respondents’ scores may be reduced, thereby making it difficult to discriminate among individuals. Additionally, the stringent standards of Guttman scaling can themselves be a disadvantage. In fact, some strategies used to overcome these difficulties can lead to additional problems. Researchers often start the development process with several items, generally selected using some rationale. Then, based on the observed results, items that weaken reproducibility are eliminated. Because these omissions often occur without theoretical backing, scale validity can be compromised.

Because of its challenges, Guttman’s technique is rarely used. As a result, the reliability and validity of this approach relative to other methods have not been established definitively.

—*Leandre Fabrigar and
Karen MacGregor*

See also Attitude Tests; Likert Scaling; Thurstone Scales

Further Reading

- Borgatta, E. F. (1955). An error ratio for scalogram analysis. *Public Opinion Quarterly*, 19, 96–100.
- Dotson, L. E., & Summers, G. F. (1970). Elaboration of Guttman scaling techniques. In G. F. Summers (Ed.), *Attitude measurement* (pp. 203–213). Chicago: Rand McNally.
- Edwards, A. L. (1957). *Techniques of attitude scale construction* (pp. 172–199). New York: Appleton-Century-Crofts.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Guttman, L. (1970). The Cornell technique for scale and intensity analysis. In G. F. Summers (Ed.), *Attitude measurement* (pp. 187–203). Chicago: Rand McNally.
- Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners* (pp. 47–51). New York: Teachers College Press.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Guest, G. (2000). Using Guttman scaling to rank wealth: Integrating quantitative and qualitative data. *Field Methods*, 12(4), 346–357.

Wealth ranking in given field sites can be problematic for a number of reasons. This article explores the usefulness of **Guttman scaling** and AnthroPac software in such contexts, using a small fishing community on the northern coast of Ecuador as an example. The author provides a step-by-step description of procedures for implementing and analyzing Guttman scale methodology and discusses the issue of construct validity. The complementary relationship between qualitative and quantitative data is highlighted throughout.

ENCYCLOPEDIA OF
MEASUREMENT

AND
STATISTICS

ENCYCLOPEDIA OF
MEASUREMENT
AND
STATISTICS

VOLUME **2**

EDITOR
NEIL J. SALKIND
UNIVERSITY OF KANSAS

MANAGING EDITOR
KRISTIN RASMUSSEN
UNIVERSITY OF KANSAS

A SAGE Reference Publication

 **SAGE Publications**
Thousand Oaks ■ London ■ New Delhi

H

Discovery consists of seeing what everybody has seen and thinking what nobody has thought.

—Albert Szent-Györgyi

HARMONIC MEAN

The *harmonic mean* is another way of expressing central tendency for a set of scores. The harmonic mean is obtained by dividing the number of observations by the sum of the reciprocal of the scores. That is, if we have n observations with scores $x_1, x_2, \dots, x_n$, the harmonic mean is computed as follows:

$$h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

There are certain situations in which the harmonic mean provides the most appropriate definition of the “average.” For example, the harmonic mean is useful when averaging rates of speed. To illustrate, suppose that a vehicle travels from city A to city B at an average speed of 40 miles per hour and returns to city A at an average speed of 60 miles per hour. The average speed is calculated as 48 miles per hour using the harmonic mean, as shown below:

$$h = \frac{2}{\left(\frac{1}{40} + \frac{1}{60}\right)} = 48.$$

In other words, the total amount of time for the round trip is equivalent to the time it would have taken to make the trip at a constant speed of 48 miles per hour. Had we used the arithmetic mean to compute the average velocity for this example, the result would have been 50 miles per hour.

The harmonic mean has applications in the behavioral sciences as well. An example can be found in the context of statistics. Suppose that two group means are compared using an independent t test. The denominator of the t statistic is the standard error,

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{pooled}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

which quantifies the sampling error associated with the mean difference in the numerator of the t formula. In this formula, s_{pooled}^2 represents the pooled variance and is computed as

$$s_{pooled}^2 = \frac{ss_1 + ss_2}{n_1 + n_2 - 2},$$

where ss_k is the sums of squares for the k th group. After some algebraic manipulation, it is possible to

rewrite the standard error formula using the harmonic mean h in place of n_1 and n_2 as follows:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s_{pooled}^2 \left(\frac{1}{h} + \frac{1}{h} \right)}.$$

To illustrate, suppose it was of interest to compute an independent t test with unequal group sizes, such as $ss_1 = 25$, $n_1 = 12$, $ss_2 = 35$, and $n_2 = 20$. Using standard formulae, the pooled variance is $s_{pooled}^2 = 2$, and the standard error of the mean difference is $s_{\bar{X}_1 - \bar{X}_2} = .52$. The harmonic mean sample size in this example is $h = 15$ (i.e., $2/[1/12][1/20]$). Substituting the harmonic mean into the formula above yields a standard error of .52—the same result obtained using unequal ns . Thus, the standard error of the mean difference using two unequal groups is identical to the standard error that would have been obtained using two groups with a common n equal to h (i.e., $n_1 = n_2 = h$).

In sum, there are certain situations in which the harmonic mean provides the appropriate definition of the “average,” as when averaging rates and when computing the average sample size from a disparate collection of ns .

—Davood Tofghi and Craig K. Enders

See also Average; Mean

Further Reading

- Hoehn, L., & Niven, I. (1985). Averages on the move. *Mathematics Magazine*, 58, 151–156.
- Kenney, J. F., & Keeping, E. S. (1962). Harmonic mean. *Mathematics of statistics, Pt. 1* (3rd ed., pp. 57–58). Princeton, NJ: Van Nostrand.

Harmonic mean article: http://en.wikipedia.org/wiki/Harmonic_mean

HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT

The Health Insurance Portability and Accountability Act (HIPAA) is considered the most significant health

care legislation since Medicare and Medicaid were created in the 1960s. It has restructured the delivery and management of health care in such a way that nearly all aspects of patient care as well as the business of health care are affected. Evaluation and assessment of such services are included in the act as well.

Signed into law by President Bill Clinton on August 21, 1996, HIPAA, as proposed by Senators Nancy Kassebaum (Republican from Kansas) and Edward Kennedy (Democrat from Massachusetts), had a twofold intent: (a) to protect the health insurance coverage of American workers and their dependents by ensuring the ability to renew or obtain health insurance in the event of a change or loss of jobs, thereby guaranteeing portability across employment settings, and (b) to reduce or eliminate discrimination against employees that is based on preexisting medical conditions. The legislation was ultimately expanded to include requirements pertaining to administrative simplification and health care abuse and fraud. The overwhelming focus of this part of the legislation pertains to privacy.

Portability

HIPAA's portability rules are intended to address a situation known as *job lock*, or individuals' reluctance to change jobs for fear of losing health coverage for preexisting conditions. HIPAA guarantees an individual's right to purchase health insurance, providing the individual (a) has had 18 months of “creditable coverage” through a group health plan, (b) does not have other health insurance and is not eligible for health coverage under another group plan (including Medicare or Medicaid), (c) did not lose health insurance for lack of payment of premiums or fraud, and (d) has exhausted coverage provided under the Consolidated Omnibus Budget Reconciliation Act of 1985. Once individuals are HIPAA-eligible, HIPAA prohibits denying them health insurance or charging them higher rates for health insurance on the basis of their health history or the health history of their dependents. HIPAA also places the following limits on exclusions due to preexisting conditions:

- In certain cases, HIPAA may limit the amount of time to which exclusions for a preexisting medical condition apply so that the diagnosis or treatment of medical conditions before this period may not be used by health plans. This *look-back period* provision defines a preexisting condition as a medical condition that has been diagnosed or treated within 6 months of the enrollment date. Thus, a medical condition for which treatment was received prior to that period would not qualify as a preexisting condition.

- In cases in which a preexisting condition is determined to exist, HIPAA limits the time during which exclusions can prevent an individual from receiving health care for that condition to 12 months.

- Exclusions for preexisting conditions can be reduced if an individual has had creditable coverage under another health plan or policy prior to enrollment in a new health plan. For example, if an employee had 6 months of creditable coverage under another group health plan, the previous coverage can be applied to the 12-month period. Thus, the new plan cannot exclude coverage for more than 6 months.

Exclusions based on preexisting conditions are not applied to newborns or adopted children who are covered under group health plans within 30 days after their birth.

Administrative Simplification

The goals of administrative simplification fall into two general categories: (a) improving the efficiency and effectiveness of health care by standardizing shared electronic information and (b) protecting the privacy and security of patient information that is stored and exchanged electronically. In addition, administrative simplification has four specific standards or rules that pertain to the handling of health data by *covered entities*, that is, health plans or institutions, health care providers, and other individuals who handle medical records and other health care information. The U.S. Department of Health and Human Services requires that these rules meet three basic criteria: they must be comprehensive, they must be suitable for any size or type of covered entity, and

they must not be negatively impacted (i.e., rendered inoperative or obsolete) by evolving technologies.

The Privacy Rule

Because patients, or their legal designees, have the ultimate authority to make health care decisions on their own behalf, HIPAA provides them with control over medical records. The Privacy Rule stipulates HIPAA's regulations with regard to organizations or individuals (i.e., covered entities) involved with *protected health information* (PHI; defined below) in any and all capacities. The Privacy Rule went into effect in April 2003 for larger entities and April 2004 for smaller ones. Already existing state privacy protections at least as stringent as HIPAA's remained in effect. Covered entities that do not comply with HIPAA's Privacy Rule risk civil or criminal monetary penalties and imprisonment.

PHI refers to any type of information (transmitted or maintained by electronic or other media) that identifies an individual with respect to the provision of, or payment for, past, present, or future physical or mental health conditions. Under HIPAA, physicians are required to ensure the privacy of patients' medical information by controlling how that information is used and disclosed. Specifically, PHI may be used or disclosed only for work-related purposes and only to the extent necessary for achieving those purposes. For example, when a patient is referred from one medical office to another, only information relevant to the patient's medical history should be conveyed (and not, for example, billing information).

Health care providers must also provide patients with a Notice of Privacy Practices, in writing, informing them of patients' rights with respect to PHI. These rights include the following:

- the right to access and copy personal medical records
- the right to request that medical information be amended
- the right to an accounting concerning individuals or institutions to which health information has been disclosed
- the right to request that access to sensitive information be limited

The Notice of Privacy Practices may also be used to inform patients of the intended use of certain information and to provide patients with the opportunity to refuse disclosure of that information. This provision applies most often to large medical facilities (e.g., hospitals) that routinely make certain information available, such as in patient directories, lists of the religious affiliations of patients, and the disclosure of certain medical information to family members or friends.

Physicians are also required to provide privacy protections on an administrative level by designating a privacy officer within their medical facility who is responsible for the training of staff and the implementation of necessary security policies and procedures. Patients must be informed of their right to complain to a facility's privacy officer in the event of problems related to the privacy of PHI and, if a complaint is not resolved, to pursue the complaint with the Department of Health and Human Services Office of Civil Rights.

Health care providers are not required to obtain explicit authorization from patients in order to disclose PHI for purposes of providing treatment, obtaining payment for treatment, or as needed as part of the operations of the provider. In addition, patient authorization is not required under circumstances when the law requires disclosure. These circumstances include compliance with public health policies, judicial proceedings, law enforcement, and emergency health care, among others. On the other hand, there are situations under which HIPAA extends extra protections regarding the disclosure of PHI. Specifically, this protection pertains to information contained in psychotherapy notes and requires separate authorization for their release. In fact, restrictions may be imposed even on patients' access to this type of information.

Providers must also have contracts with any individuals or companies who conduct business on behalf of the provider if that business entails access to PHI. These contracts specify compliance with HIPAA's privacy rule, particularly with regard to the use or disclosure of PHI.

The Privacy Rule also protects health information when used for research purposes. Guidelines stipulate how researchers can access and subsequently use that information. For example, in addition to requiring

researchers to obtain informed consent in order to use PHI from current patients, the rule also applies to previously collected health care data that may exist in clinical records or registries. Investigators cannot use these data without first obtaining consent from former patients who are still alive.

The Security Rule

This rule, which overlaps in some ways with components of the Privacy Rule, took effect in April 2005 for larger entities. It stipulates the minimum security standards for PHI that is stored or transmitted electronically, including by e-mail, by requiring that safeguards be in place in order to ensure the integrity and confidentiality of health information. Although these regulations do not mandate encryption in order to secure PHI that is electronically stored or transmitted, most covered entities do encrypt this information in some way in order to be in compliance with HIPAA regulations.

All safeguards must incorporate three major components: (a) the assessment of potential risk to and vulnerability of electronic PHI; (b) the development, implementation, and maintenance of necessary security measures to reduce or eliminate risk; and (c) the documentation thereof. Although HIPAA specifies the safeguards for protecting PHI, covered entities are left to decide how those safeguards will be implemented. These decisions take into account the types of services an entity provides and the types of technology it uses, as well as its operational needs and its resources.

Administrative Safeguards

Administrative safeguards incorporated into HIPAA require that covered entities adopt written procedures that identify employees who will have access to PHI and designate a privacy officer to implement those procedures. In addition, covered entities must have procedures in place for both routine and emergency data backup and recovery. Finally, there should be policies and procedures for routine internal audits, the purpose of which is to review potential violations of security.

Physical Safeguards

Additional safeguards are meant to prevent unauthorized access to protected information. They include controls over the installation and removal of software and hardware and the careful placement of workstations and patient files in high-traffic public areas in order to reduce unauthorized access to protected information.

Technical Safeguards

Technical safeguards provide controls over access to computer systems and the security of electronic transmission of PHI over networks. Covered entities must ensure the integrity of data stored in their systems and provide the government with evidence of compliance with these HIPAA provisions.

The Identifier Rule

This rule proposes a standard for identifying health care providers and for improving the efficiency of transmitting health information electronically. Under this rule, all providers receive a national provider identifier, which is a number that is unique to a particular health care provider or insurer and which must then be used in all electronic transactions identified in the HIPAA regulations. This number may also be used as a way of identifying a provider in external correspondence, on prescriptions, in internal communications, and so on. With a set of national identification numbers in place, processes such as eligibility determination and claims processing may improve. In addition, it is believed that the national provider identifier will help identify the fraudulent use of multiple identifiers by both individuals and providers.

The Transaction and Code Sets Rule

The Transaction and Code Sets Rule took effect in October 2003 for all covered entities. Like the Security Rule, it pertains only to PHI in its electronic form. HIPAA authorized the Department of Health and Human Services to mandate use of Electronic Data Interchange (EDI) in order to standardize medical records and to simplify the use of electronic

transactions for claims, remittances, eligibility verifications, referrals, and such. In addition, the Department of Health and Human Services was permitted to identify organizations called *designated standards maintenance organizations* to develop, maintain, and modify EDI standards. In general, EDI applies to the exchange of data between or among computer systems. A subcommittee of the American National Standards Institute provides the standardized data format for health insurance transactions (e.g., claims, payment, eligibility). Within these transactions, medical diagnoses are coded according to the International Classification of Diseases, and procedures are categorized according to coding systems, such as Current Procedural Terminology, Healthcare Common Procedure Coding System, and the Code on Dental Procedures and Nomenclature. It is expected that the Transaction and Code Sets Rule will result in the anticipated efficiency and savings of administrative simplification.

HIPAA is important health care legislation meant to ensure continued health coverage as well as the privacy and security of personal health information. However, the practical and financial impact of HIPAA on the practice of medicine and clinical research has been considerable and remains a matter of contention between the medical community and the government.

—Carole E. Gelfer

See also Measurement

Further Reading

National standards for health care: <http://www.hhs.gov/ocr/hipaa/>

U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services. HIPAA—General Information: <http://www.cms.hhs.gov/hipaa/hipaa2/>

U.S. Department of Health and Human Services. Summary of the HIPAA Privacy Rule: <http://www.hhs.gov/ocr/privacy/summary.pdf>

HELLO-GOODBYE EFFECT

The Hello-Goodbye Effect refers to the bias caused by patients who exaggerate their problems before a

treatment (or any intervention), hoping to be eligible for the therapy, and minimize their problems at the end of the treatment, hoping to please the therapist. The Effect has two components. The first is related to the Faking Bad Bias, whereby participants try to appear sick to qualify for support. The second is related to the Faking Good Bias (or Social Desirability Bias or Obsequiousness Bias), whereby participants may systematically respond in the direction they perceive to be desired by the investigator.

An example of the Hello-Goodbye Effect is found in a review of pain measures of patients before and after acupuncture treatment. Pain is the most frequent complaint of patients referred for acupuncture. However, it is difficult to measure because it cannot be measured directly and must be measured on the basis of the patients' response. For example, the patient may be asked to select from five pain categories ranging from *none* to *intense* or from a numerical rating scale ranging from 0 to 10. Results of the measurement can be greatly influenced by the individual's genotype, culture, conditioning, education, and so on. The review suggested that, when patients first present for treatment, they need to justify their request for help, so there is a subconscious tendency to exaggerate symptoms. After patients have received treatment, they may want to please the therapist, or at least not hurt the therapist's feelings, so there is a tendency to minimize symptoms.

The Hello-Goodbye Effect may result in surprising improvements in the symptoms after a treatment. In extreme cases, even totally ineffective treatments can produce an improvement in treatment results.

In clinical practice, it is difficult to prevent or minimize the Hello-Goodbye Effect. But in research, it can be minimized by letting participants know that the information they provide (as recorded on the forms or questionnaires) will not be seen by the therapist.

One must be careful not to confuse the Hello-Goodbye Effect with other, distinctly different effects, which in some cases may produce similar outcomes. These effects may include Apprehension Bias (certain measures may alter systematically from their usual levels when the participant is apprehensive; e.g., blood pressure may change during medical interviews); Attention Bias (or the Hawthorne Effect;

study participants may systematically alter their behavior when they know they are being observed); Culture Bias (participants' responses may differ because of culture differences); and the placebo effect (measurable, observable, or felt improvement in health attributable not to treatment but to a placebo, which is a medication or treatment believed by the therapist to be inert or innocuous).

—Bernard C. K. Choi and Anita W. P. Pak

See also Measurement

Further Reading

- Aday, L. A. (1996). *Designing and conducting health surveys* (2nd ed.). San Francisco: Jossey-Bass.
- Baines, L. S., Joseph, J. T., & Jindal, R. M. (2004). Prospective randomized study of individual and group psychotherapy versus controls in recipients of renal transplants. *Kidney International*, 65, 1523–1755.
- Carroll, R. T. (2003). *The skeptic's dictionary: A collection of strange beliefs, amusing deceptions, and dangerous delusions*. Chichester, UK: Wiley. Retrieved August 7, 2005, from <http://www.skeptdic.com/placebo.html>
- Choi, B. C. K., & Pak, A. W. P. (1998). Bias, overview. In P. Armitage & P. Colton (Eds.), *Encyclopedia of biostatistics: Vol. 1* (pp. 331–338). Chichester, UK: Wiley.
- Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. *Preventing Chronic Diseases*, 2, 1–20. Retrieved August 6, 2005, from http://www.cdc.gov/pcd/issues/2005/jan/04_0050.htm
- Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use*. Oxford, UK: Oxford University Press.
- White, A. (1998). Measuring pain. *Acupuncture in Medicine*, 16, 1–11. Retrieved August 6, 2005, from <http://www.medical-acupuncture.co.uk/journal/1998nov/six.shtml>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Walters, G. D., Trgovac, M., Rychlec, M., Di Fazio, R., & Olson, J. R. (2002). Assessing change with the Psychological Inventory of Criminal Thinking Styles: A controlled analysis and multisite cross-validation. *Criminal Justice and Behavior*, 29(3), 308–331.

Lee Cronbach, a famous measurement specialist, coined the term **Hello-Goodbye Effect** to explain how participants in programs create an overly positive image of the effects of a program

when exiting that program. In this study, the sensitivity of the Psychological Inventory of Criminal Thinking Styles (PICTS) to psychotherapeutically assisted change was evaluated in a series of three studies. In the first study, a repeated measures ANOVA revealed significant group (participant-waiting list) and time (pretest-posttest) effects, and a paired t test indicated significant reductions on the PICTS Current Criminal Thinking scale in group participants. In the second study, inmates from four different settings who were participating in programs of differing length and content achieved statistically significant temporal reductions on the Current and Historical Criminal Thinking scales. Results from the third study showed that significant pretest-posttest reductions on the Current Criminal Thinking scale were specific to good-prognosis participants. There is also some reason to believe that the Hello-Goodbye Effect may have had a stronger impact on those who participated in the shorter group.

HETEROSCEDASTICITY AND HOMOSCEDASTICITY

Homoscedasticity and heteroscedasticity refer, respectively, to whether the variances of the predictions determined by regression remain constant or differ. Heteroscedasticity is perhaps most often considered in cases of linear regression through the origin, although that is by no means the limitation of its usefulness. This is, however, a good example with which to explain the concept. Consider a linear regression to predict tree height for a certain species of tree as a function of a single regressor: tree diameter at the trunk. When the tree trunk diameter is nearly zero, the height of that tree is nearly zero. Thus we could consider linear regression through the origin. Predictions of tree height, y , say $\hat{y}$, using the tree trunk diameter, x , as a single regressor, may vary greatly in accuracy as x and y become larger. This is reasonable because it would not be expected that at a given confidence level, one would have predicted tree heights of, say, 1.8 meters ± 0.5 meter, 9.7 meters ± 0.5 meter, and 17.6 meters ± 0.5 meter. It would be more logical to expect that the smaller values of $\hat{y}$

would have smaller variances (thus smaller standard errors), and the larger predictions would have larger variances.

Similarly, a factory that produces y “widgets” a day while employing x workers would be expected to have larger variances in predictions for larger y with larger x than would be the case for smaller factories. Variance could be explained due to differences in factory equipment type, age, efficiency and varying policies, as well as the quality of the widgets produced. It is intuitive that when there are zero employees, zero widgets will be produced, and that larger predictions of widgets will have larger variances. Let us once again consider linear regression through the origin and describe this with the following equation: $y_i = bx_i + e_i$, where $e_i = e_{0i}x_i^\gamma$. Note that as x becomes larger, the estimated residuals, e_i , would generally be larger, and thus the variance is larger. Let e_{0i} be the random factor of the residual. (More than one regressor could be involved, but this possibility is not shown here.)

When we assume that variance is constant and that the regression is not necessarily through the origin, linear regression in that case is referred to as ordinary least squares (OLS) regression. Linear regression that allows for variance that is not constant is called weighted least squares (WLS) regression. The weights are not the same as design-based sample weights used in survey estimation (the inverse of the probabilities of selection) and should not be confused with them when dealing with surveys. (Linear regressions are often used in model-based approaches to survey statistics.) The weights at hand are instead those that indicate how much confidence we have in the various data points when using them to estimate a regression line. Mathematically, the weights in the above equation are $w_i = x_i^{-2\gamma}$ (not derived here but shown in many econometrics and statistics books and other resources). When $\gamma = 0$, the weights are all equal (OLS regression). When $\gamma = 0.5$, with regression through the origin, this case provides us with the classical ratio estimate (CRE), such that

$$b = \frac{\sum_{i=1}^n w_i x_i y_i}{\sum_{i=1}^n w_i x_i^2} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

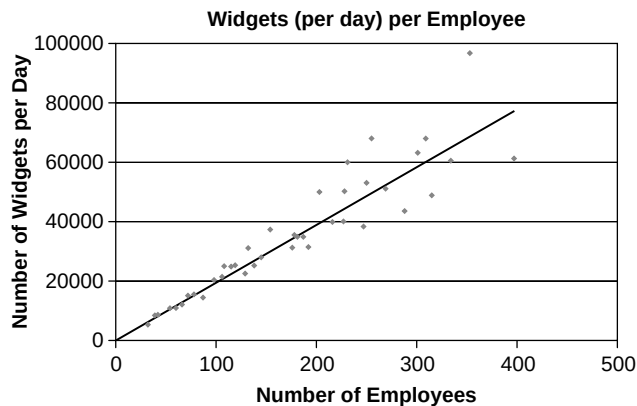


Figure 1 A Heteroscedastic Linear Regression Relationship Between Widgets Produced and Workforce Size

Many practitioners and academicians seem to avoid WLS regression, or adjustments for it, because they know assumptions must be made about the regression weights. However, OLS regression assumes constant variance, which also assumes something about regression weights. In OLS regression, regression weights are all equal, which is often a very bad assumption. Thus, choosing to ignore what the weights might be is to assume they are all equal, which might be a decision one would not make consciously. There are methods to estimate the value of gamma above, based on the data available, but often the CRE or another specific WLS estimate may be robust (i.e., flexible or generally accurate) for a practitioner's purposes.

—James Randolph Knaub, Jr.

See also Variance

Further Reading

- Abdi, H. (2004). Least squares. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage. (Note on page 2 that ϵ is the sum of squared residuals, not just a residual or "error," as stated.) Retrieved April 11, 2006, from <http://www.utdallas.edu/~herve/Abdi-Least-Squares-pretty.pdf>
- Brewer, K. R. W. (2002). *Combined survey sampling inference: Weighing Basu's elephants*. London: Oxford University Press/Arnold.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman & Hall/CRC.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Griffiths, W. E., Hill, R. C., & Judge, G. G. (1993). *Learning and practicing econometrics*. New York: Wiley.

Knaub, J. R., Jr. (1997). *Weighting in regression for use in survey methodology*. Retrieved April 11, 2006, from <http://interstat.statjournals.net/>

Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Sweet, E. M., & Sigman, R. S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association. (Note that section 2.6 references three methods for estimating γ .)

HIERARCHICAL LINEAR MODELING

Researchers across various disciplines often deal with data that have multilevel, hierarchical, or nested structure. Methodologists from various disciplines refer to the statistical methods for analyzing hierarchically structured data differently. Researchers in behavioral and social sciences refer to them as hierarchical linear modeling (HLM) or multilevel modeling. In statistics, the process is referred as covariance components modeling. In biometrics, it is referred as random-effects modeling and mixed-effects modeling, and in econometrics, it is referred to as random-coefficient regression modeling.

These hierarchical models explicitly model lower and higher levels in the hierarchy by taking into consideration the interdependence of individuals within the groups. For example, in a two-level HLM analysis, the emphasis is on how to model the effects of explanatory variables (predictors) at one level on the relationships occurring at another level. These multilevel structured data present analytical challenges that cannot be handled by traditional linear regression methods, because there is a regression model for each level.

Applications of HLM

HLM is usually applied to different kinds of hierarchically structured data. One of the most common applications involves multilevel data with continuous outcomes. In such data, individuals have continuous outcomes, and they are clustered (nested) within groups or organizations. An example of such basic two-level data would be students nested within classrooms, with math achievement as the outcome of interest. Multilevel models can have more than two levels; students nested within classrooms, which are nested within schools, would be an example of a three-level model. A second application involves modeling hierarchically structured data with other types of outcomes, such as binary responses, count data, or multinomial data. A third application involves modeling growth-structured data for which longitudinal measurements on each individual constitute the first level and individuals constitute the second level in a two-level growth model. A logical extension of this application to a three-level growth model is individuals nested within organizations. A fourth application involves measurement models wherein responses to a set of items are nested within individuals. A fifth application is multilevel modeling of meta-analytic data sets, with variations in the effect sizes from collected primary studies modeled by study and sample characteristics. Applications of HLM can be extended beyond those listed above depending on an appropriate conceptualization of the investigated problem within the HLM structure. Given the potential complexity of multilevel modeling and its many analytical applications, with unique hierarchical linear model conceptualizations for each application, a two-level hierarchical linear model is presented below.

Two-Level Full Hierarchical Linear Model

The full two-level hierarchical linear model is characterized by two levels, individuals nested within groups or organizations, and measured predictors for each level. In this full hierarchical linear model, researchers

are primarily interested in assessing the effects of both individuals' characteristics and groups' characteristics on the outcome (e.g., performance, achievement, or any other continuous outcome).

Level-1 (Individual-Level) Model

The individual-level model specifies the relationships between various individual characteristics, as independent explanatory variables (predictors), X_{ij} , and the outcome (dependent) variable, Y_{ij} . Given that we have p predictors for n_j individuals within j groups, the Level-1 model takes the form

$$Y_{ij} = \beta_{0j} + \beta_{pj} X_{pij} + r_{ij}, \quad (1)$$

where

i equals $1, 2, 3, \dots, n_j$ individuals within group j ,

j equals $1, 2, 3, \dots, J$ groups,

β_{0j} represents the intercept for the Level-1 model,

β_{pj} represents p regression coefficients capturing the effect of the p predictors X_{pij} on the outcome, and

r_{ij} is a random error assumed to be normally distributed with mean zero and a common variance, σ^2 .

Thus, the Level-1 model yields j separate sets of regression estimates for the intercept and the p slopes. These regression estimates are then modeled as outcomes in Level-2 models.

Level-2 (Group-Level) Model

The j estimates for each of the regression coefficients (intercept and the p slopes) from Level 1 become an outcome in the Level-2 model, which can be modeled by Level-2 characteristics (predictors). This Level-2 regression model takes the forms

$$\beta_{0j} = \gamma_{00} + \gamma_{0q} G_{qj} + U_{0j} \quad (2)$$

$$\beta_{pj} = \gamma_{p0} + \gamma_{pq} G_{qj} + U_{pj}, \quad (3)$$

where

G_{qj} are Level-2 predictors (e.g., organization size, organization type),

γ_{00} , γ_{0q} , γ_{p0} , and γ_{pq} are Level-2 fixed effects regression coefficients, and

U_{0j} and U_{pj} are the Level-2 residuals that have a multivariate normal distribution with a zero mean vector and variance-covariance matrix T with τ_{00} , τ_{pp} , τ_{0p} , and $\tau_{p'p}$ components where $p' \neq p$.

No-Predictors Two-Level Hierarchical Linear Model

A much simpler model than the full hierarchical linear model presented above is a two-level no-predictors hierarchical linear model, in which there are no explanatory variables in either level of the model. This no-predictors model is referred to in the literature as a one-way random effects analysis of variance model, unconditional model, null model, or empty model.

No-Predictors Individual-Level (Level-1) Model

A no-predictors individual-level model is represented as

$$Y_{ij} = \beta_{0j} + r_{ij}, \quad (4)$$

where

Y_{ij} is the outcome for individual i ,

$i = 1, 2, \dots, n_j$ in group j ,

$j = 1, 2, \dots, J$,

β_{0j} is the mean of the outcome in group j , and

r_{ij} is Level-1 random error with a mean of zero and a variance of σ^2 .

No-Predictors Group-Level (Level-2) Model

At the second level (group level), β_{0j} is allowed to vary randomly across groups, with a mean γ_{00} and error term U_{0j} as follows:

$$\beta_{0j} = \gamma_{00} + U_{0j}, \quad (5)$$

where

γ_{00} is the overall grand mean of the outcome across all schools, and

U_{0j} is the error term representing the deviation of the each organization's mean outcome from the grand mean, γ_{00} .

The variance of U_{0j} is τ_{00} .

Amount of Variance (R^2) Explained

The variance estimates from the no-predictors and a particular full hierarchical linear model provide estimates of R^2 for the Level-1 and Level-2 models. R^2 for the Level-1 model represents the percentage of variance in the Level-1 outcome accounted for by the Level-1 predictors, and it is equal to

$$R^2_{\text{Level-1 Model}} = \frac{\sigma^2_{\text{No-predictors model}} - \sigma^2_{\text{Full model}}}{\sigma^2_{\text{No-predictors model}}}. \quad (6)$$

There are $(p + 1)$ estimates of R^2 for the Level-2 model, one for the Level-1 intercept and one for each of the p predictors used to predict the outcome in Level 1. R^2 for the Level-1 intercept represents the percentage of the intercept variance accounted for by the Level-2 predictors, and it is equal to

$$R^2_{\text{Level-2 intercept model}} = \frac{\tau_{00} - \text{No predictors in Level-2 model} - \tau_{00} - \text{Full Level-2 model}}{\tau_{00} - \text{No predictors in Level-2 model}}. \quad (7)$$

Each of the p R^2 s for Level-1 slopes represents the percentage of a particular p slope variance accounted for by the q Level-2 predictors, and it is equal to

$$R^2_{\text{Level-2 slope model}} = \frac{\tau_{pp} - \text{No predictors in Level-2 model} - \tau_{pp} - \text{Full Level-2 model}}{\tau_{pp} - \text{No predictors in Level-2 model}}. \quad (8)$$

Hierarchical Linear Model Estimation and Statistical Testing

Generally, hierarchical linear model analyses provide three kinds of estimates of parameters. First are

estimates for the Level-2 fixed effect parameters, γ (Equations 2, 3, and 5). HLM provides a t test for testing the hypothesis that each of these parameters is significantly different from zero. The second are empirical Bayes estimates of Level-1 regression coefficients, β s. The third are Level-1 residuals variance (σ^2) and the variances-covariances of the Level-2 residuals (τ_{00} representing the variance of U_{0j} , τ_{pp} representing the variance of U_{pj} , τ_{0p} representing the covariance between U_{0j} and U_{pj} , and $\tau_{p'p}$ representing a covariance between $U_{p'j}$ and U_{pj} where $p' \neq p$). A chi-square test statistic tests the null hypothesis that a particular Level-2 residual variance (τ_{00} or τ_{pp}) is significantly different from zero.

Assumptions of the Hierarchical Linear Model

As any statistical method, HLM analysis requires several statistical assumptions to ensure the validity of estimating and testing the fixed and random effects' coefficients. The assumptions for a two-level hierarchical linear model are as follows:

Level-1 residuals and predictors are uncorrelated.

Level-1 residuals are independent and normally distributed with mean zero and common variances σ^2 .

Level-2 residuals are independent, multivariate, and normally distributed, with a mean vector of zero and a variance-covariance matrix T .

Level-2 residuals and predictors are uncorrelated.

Level-1 and Level-2 residuals are uncorrelated.

Intraclass Correlation Coefficient

One advantage of doing HLM analysis is to accommodate the violation of the assumption of independent observations. Intraclass correlation assesses the degree of violating such an assumption. It measures the degree of dependency among individuals within groups due to common experiences of individuals within groups. It is the ratio of the Level-2 variance for an unconditional model, τ_{00} , to the total variance ($\tau_{00} + \sigma^2$):

$$ICC = \tau_{00} / (\tau_{00} + \sigma^2). \quad (9)$$

Centering Level-1 Predictors

In a traditional multiple regression model, the intercept represents the expected value of the outcome measure when all predictors have zero values. There are situations in which some predictors do not have meaningful zero value. For example, it does not make sense to say that a person has zero intelligence or achievement. Thus, centering Level-1 predictors becomes necessary in order to have an interpretable and meaningful intercept. Two kinds of centering are available in hierarchical linear models:

1. *Grand mean centering*, in which all values of a particular predictor, X_{ij} , are centered on its grand mean, $\bar{X}$, across all j groups as follows:

$$X_{ij} - \bar{X}. \quad (10)$$

2. *Groups mean centering*, in which all values of a particular predictor, X_{ij} , are centered on its group mean, $\bar{X}_j$, as follows:

$$X_{ij} - \bar{X}_j. \quad (11)$$

HLM Software Packages

Advancements in statistical methods and computer technologies within the past two decades have led to many specialized software packages for analyzing hierarchically structured data. Examples include HLM, MLnWin, VARCL, SAS, SPSS, Proc Mixed, MIXOR, and MIXREG.

—Sema A. Kalaian and Rafa M. Kasim

See also Multivariate Analysis of Variance (MANOVA)

Further Reading

- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

HIGH-STAKES TESTS

Standardized achievement tests carry high stakes if serious consequences are attached to student

performance on them. High-stakes tests are typically given in the three core academic subjects, reading, writing, and mathematics, and serious consequences are attached to individual student performance or student performance averaged at the classroom (teacher), school, or school district level.

For individual students, high test scores bring college scholarships and other forms of recognition, including academic awards and marks of distinguished achievement on high school diplomas. Low test scores bring retention in grade at the elementary level or denial of a high school diploma at the secondary level. If a third-grade student does not pass a third-grade test assessing reading acuity, the student could be retained in (i.e., expected to repeat) third grade. These tests are known as grade-to-grade promotion exams. If a high school student does not pass a high school test assessing mathematics acuity, a high school diploma could be denied until the student passes the math test. These tests are known as high school graduation or exit exams and, to date, are the most frequently used form of high-stakes tests in the country.

Students' high test scores bring teachers financial bonuses, salary increases, and more professional freedoms (e.g., choices in classes taught or scheduling preferences). Low test scores may cause a teacher to be fired or transferred to a different school and may be assumed to indicate that the teacher needs to undergo professional development to become better equipped to improve student performance.

High test scores merit monetary awards for local administrators, schools, and school districts. They also draw public acclaim for schools, districts, and school administrators when scores are published in local newspapers. Conversely, low test scores bring public scrutiny. They may be used as an indication that a school administrator should be terminated or that a school or school district should be reconstituted; taken over by a state, private, or charter school entity; no longer accredited; or simply closed. If students attend schools with a history of low test scores, the students may choose to transfer to schools with higher test scores.

High-Stakes Tests in Theory

The aforementioned stakes have been attached to tests by educational policymakers who believe that attaching incentives to learning and sanctions to poor performance will increase academic achievement. It is their belief that raising academic standards and attaching serious consequences to tests that measure the extent to which students meet these standards will inspire students to learn more and motivate teachers and administrators to implement more effective educational programs.

It is also their belief that these measures will help states target the students, teachers, schools, and school districts that are most and least successful—success being defined by student performance in reading, writing, and mathematics on high-stakes tests. The most successful schools may be examined and their practices replicated to help the least successful schools meet higher standards. Students in the least successful schools may be targeted for special programs to help them achieve higher standards. Theoretically, such efforts will help otherwise failing students pass and may reduce the achievement gap between traditionally marginalized students—students from racial minority, non-English-speaking, and economically disadvantaged families—and their more affluent, predominantly White peers.

A Brief History of High-Stakes Tests

In the 1970s, many feared that students in the United States were not matching the levels of performance posted by students in other industrialized nations—nations with which the United States was in global competition. A general sense of educational mediocrity was blamed for jeopardizing the United States' global superiority. The general public was dissatisfied with the education system and disappointed in most reform efforts, which failed to demonstrate significant improvements in student achievement.

Florida implemented the first high-stakes testing policy known to educational reformers. Florida implemented a high school graduation exam that students had to pass in order to receive a high school diploma.

The test included very basic skills (also known as minimum competency skills). A few other states followed Florida's lead, implementing minimum competency tests of the same sort.

In the 1980s, the National Commission on Education, in its publication *A Nation at Risk*, called for an end to minimum competency testing. Tests that ensured minimum competency levels of learning were accused of promoting a low level of achievement and a diluted set of curriculum standards. According to the Commission, improving the condition of education in the United States should mean holding students accountable for meeting higher standards, to be set forth by state departments of education.

In the 1990s, the standards-based movement arose. Students were increasingly held accountable for meeting higher, more rigorous curriculum standards. A flood of legislation mandating standardized testing reforms spurred more states to rely on what we now know as high-stakes tests.

By 2002, more than half the nation had implemented or had plans to implement high-stakes testing policies when President George W. Bush signed the No Child Left Behind (NCLB) Act into law. NCLB requires that all states implement stronger accountability policies to ensure that elementary students in Grades 3–8 and high school students meet higher academic standards by the year 2014. Under NCLB, states were forced to execute stronger accountability policies or lose federal education funds.

Benefits of High-Stakes Tests

Some research studies have provided evidence that high-stakes tests have increased student achievement. Other research studies have presented counterevidence showing that states that have implemented high-stakes testing policies have fared no better in terms of student achievement than other states that have not yet implemented such policies. High-stakes tests have not done much to increase academic achievement and, if anything, might be increasing the achievement gap between White students and their less affluent, minority peers. This, the primary question with which most educators

are concerned, causes the fiercest debates. Researchers have yet to agree on what effects high-stakes tests have had on increasing academic achievement.

Perhaps the greatest benefit associated with high-stakes testing policies is that states have revisited, revised, and raised their standards to meet what professional teacher organizations (e.g., the National Council of Teachers of Mathematics) regard as essential subject knowledge. Because teachers are being held accountable for teaching the content included in these standards, more consistency across subjects and better uniformity across classrooms, schools, and even state borders has ensued. However, not everyone in the greater education profession agrees that this consistency is a benefit. Along with an increase in standardization has come a decrease in the professional autonomy a teacher has in the classroom.

Another benefit of implementing high-stakes testing policies is that monies are being targeted toward students most in need of help. Remediation programs have been developed to help students who fail high-stakes tests learn the material necessary to pass each test. Although teaching students how to pass high-stakes tests is of questionable utility and principle, these remediation efforts are being targeted toward students who went without them before NCLB.

A final benefit may be that student test scores can be used for diagnostic purposes, helping teachers better understand student comprehension and identify students for individualized instruction. However, in most states, high-stakes tests are administered in the spring, and the test results are returned during the summer, after school is out. In the fall, students start a new school year with a new teacher. To date, the diagnostic benefits of high-stakes tests have gone unrealized and, ironically, may be attained only as greater numbers of students are held back in grade.

Negative Effects of High-Stakes Tests

Most students who are subjected to the negative effects of high-stakes tests are students from traditionally marginalized backgrounds. These students have the most to lose if they fail high-stakes tests, and they cause the

most worry when teachers, administrators, and schools are held accountable for their substandard results. Consequently, teachers of these students and some administrators have employed a multitude of questionable test preparation practices to help their schools and students perform well on high-stakes tests.

Teaching to the test occurs when teachers teach students only those things the teachers know will be on the high-stakes test. A teacher who has administered a few high-stakes tests may gain some understanding of the test content and teach students only those concepts likely to be tested. A teacher may rehearse students for a high-stakes test with clone items, which look exactly like the items on previous forms of the high-stakes test but with the names and the numbers changed. A teacher may also copy an actual test and use it to drill students for the upcoming test.

Narrowing of the curriculum occurs when teachers teach less than they are supposed to, omitting what will not be included on the high-stakes test. State standards may state that a tenth-grade mathematics teacher must teach graphing of equalities and inequalities. A teacher aware that questions requiring students to graph inequalities are never included on the high school graduation exam might simply omit this lesson in order to concentrate more effort on graphing equalities instead. Two months before high-stakes tests are administered, a principal may omit recess, art, and music and replace science with mathematics and social studies with language arts to intensify math, reading, and writing instruction.

Traditionally marginalized students are also subjected to creative exclusion and exemption practices. Students with histories of poor academic performance may be encouraged to stay home and miss high-stakes tests or might be suspended or expelled before high-stakes tests are administered. Students may be exempted from high-stakes testing by being designated English language learners even though they speak English fluently enough to participate. Students may be purposely designated severely handicapped when by law their handicap should not prevent them from taking the test. School personnel would rather these students not take part in high-stakes tests for fear their scores would bring down the school's average scores, placing the school, the administrators, and the teachers in jeopardy.

In addition, faced with the increasing pressures associated with high-stakes tests, traditionally marginalized students are likely to leave school early. Researchers have shown that these students are more likely than other students to drop out of school after failing high-stakes tests, only to enter the workforce or an alternative high school diploma program.

High-stakes tests have also been shown to have a negative impact on teachers and administrators. Countless newspaper articles have described ways teachers and administrators have cheated on high-stakes tests. A teacher may allow students more time than mandated to complete a high-stakes test; may walk around the classroom providing students with hints, definitions, or answers; may tell students to rethink particular questions; and may manually change students' answers. There is also evidence that teachers are leaving the grades in which high-stakes tests are given in favor of grade levels not yet subjected to the tests. Other teachers may be leaving public schools to teach in the private sector.

Administrators may brief teachers on what will be included on an upcoming high-stakes test; may make copies of secure tests and distribute them to teachers before the official high-stakes test is administered; or may change certain students' identification numbers, making their score sheets invalid and therefore excluded from the school's composite statistics. Administrators may hire consultants who encourage teachers to focus instruction only on those students who the teachers feel have a fighting chance of passing the high-stakes test. Administrators may use funds, oftentimes entire textbook budgets, to purchase test preparation booklets filled with test practice worksheets guaranteed to boost test scores provided students rehearse one activity after the other. Administrators may also concentrate all their efforts in one subject area—and celebrate significant gains in test scores—only to realize that significant losses in the other subjects were posted simultaneously.

All these tactics result in spurious test score gains unrelated to true gains in learning. So when considering whether high-stakes tests help students meet higher standards, one must consider the factors related to true gains in academic achievement. The validity of test score gains is increasingly

compromised as serious consequences are increasingly attached to tests.

—Audrey Amrein-Beardsley

See also Measurement; Reliability Theory; Validity Theory

Further Reading

- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). Retrieved July 18, 2005, from <http://epaa.asu.edu/epaa/v10n18>
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.
- Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation Press.

Stronger Accountability component of NCLB: <http://www.ed.gov/nclb/landing.jhtml>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? *Theory and Research in Education*, 2(3), 283–308.

Many significant changes in perspective have to take place before efforts to learn the content and capabilities of children's minds can hold much sway in educational testing. The language of testing, especially of **high-stakes testing**, remains firmly in the realm of "behaviors," "performance," and "competency," defined in terms of behaviors, test items, or observations. What is on children's minds is not taken into account as integral to the test design and interpretation process. The point of this article is that behaviorist-based validation models are ill founded and that tests should be based on cognitive models that theorize the content and capabilities of children's minds in terms of such features as metacognition, reasoning strategies, and principles of sound thinking. This approach is the one most likely to yield the construct validity for tests long endorsed by many testing theorists. The implications of adopting a cognitive basis for testing, which might upset many current practices, are explored.

HISTOGRAM

A histogram is nothing more than a graphical representation, or picture, of a set of data. In that sense, it is like a pie chart. However, while most pie charts show the relative size of various categories of a qualitative variable (e.g., favorite flavor of ice cream), a histogram provides a picture of data that are quantitative in nature (e.g., examinees' earned scores on an examination).

A histogram reveals three things about a set of scores: the place(s) where scores tend to congregate along the score continuum, the degree to which the scores are spread out, and the possibility that the full set of scores can be referred to by a term such as *normal* or *skewed* or *rectangular*. Stated differently, three important questions can be answered (at least in an approximate fashion) by quickly glancing at a histogram:

1. What is the *average* score?
2. How much *variability* is there among the scores?
3. What kind of *shape* does the distribution have?

Example

Figure 1 contains an example of a histogram. This histogram was created to show the published weight of players on a professional football team. In this histogram, the abscissa (i.e., the baseline) corresponds to the variable of interest, player weight, and the ordinate (i.e., the vertical axis, on the left) represents the frequency of the team's players who have a given weight. There are nine bars in this histogram, and the height of each bar indicates how many football players were in the weight interval indicated on the abscissa beneath the bar. Thus, this histogram shows that 3 of the team's players weighed somewhere between 160 and 179 pounds that 11 of the players weighed between 180 and 199 pounds, and so on.

A quick glance at Figure 1 allows us to answer the three questions concerning average, variability, and shape. More players were in the 200- to 219-pound interval than in any other interval, so that interval is the modal interval. The weight intervals, in combination

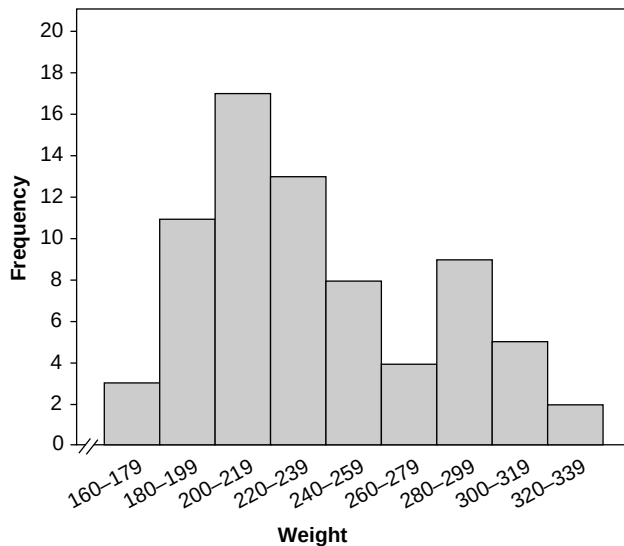


Figure 1 Example of a Histogram

with the heights of the bars, make it clear that there is a great deal of variability among the players' weights. Finally, the shape of the distribution appears to be skewed to the right (i.e., positively skewed).

Histograms Versus Bar Graphs

Although a histogram resembles a bar graph (mainly because both are made up of bars), there is one important difference between these two kinds of graphical representations. Like a pie chart, a bar graph displays data that correspond to a qualitative variable. A histogram, in contrast, displays data corresponding to a quantitative variable. This difference has an important influence on the ordering of the bars that make up a histogram or a bar graph. With bar graphs, the ordering of the bars is fully arbitrary. With histograms, the positioning of the bars cannot be scrambled; instead, they must follow (from left to right) the small-to-large feature of the baseline's variable. Because the ordering of bars is arbitrary in a bar graph, the concept of distributional shape does not apply to a bar graph.

Different Kinds of Histograms

There are many types of histograms. The two most common types are simply pictures of frequency distributions. The histogram shown in Figure 1 has intervals on the abscissa, so it corresponds to a grouped frequency

distribution. A slightly different kind of histogram has a separate bar over each individual score along the continuum of earned scores. That kind of histogram corresponds to an ungrouped frequency distribution.

Other kinds of histograms can be created. Many histograms have the vertical axis marked off so as to indicate proportions (or probabilities) rather than frequencies. Sometimes (but not often), the scores or intervals of scores are put on the vertical axis, with the horizontal axis marked off to indicate frequencies or proportions; here, the bars extend horizontally to the right from the ordinate. Occasionally, bars will be replaced by thin lines (because bar width in a histogram has no meaning).

Sometimes, two histograms will be merged. For example, a histogram for males and a histogram for females could be put into a single histogram, and there are two ways to do this. The more frequently seen method for combining histograms is to have two bars—one for each of the separate groups—over each score or score interval (with different colors or patterns indicating which bar corresponds to which group). A second way to merge two histograms is to create what's called a *bihistogram*. This kind of histogram has bars going up from the abscissa for one of the two groups and bars extending down from the abscissa for the other group.

Warnings

It is appropriate to offer three “warnings” to those who create or interpret histograms. First, remember that, in histograms like that shown in Figure 1, varying the interval width can alter the shape of the histogram. Second, realize that the difference among the bars' heights can be exaggerated if the vertical axis is “cut.” (This would happen, for example, if every bar had a frequency greater than 75 and if the ordinate therefore was set up to show only those values between 75 and 100.)

Finally, it should be noted that there are more precise ways to answer the queries that one can answer by looking at a histogram. By putting gathered scores into any of several existing formulas, it is possible to provide numerical measures of the average score (e.g., by computing the arithmetic mean or median), the amount of variability (e.g., by computing the

standard deviation or interquartile range), and the distribution's shape (e.g., by computing indices of skewness and kurtosis). Whereas the advantage of a histogram is its ability to provide quick answers to the questions about average, variability, and shape, the histogram's limitation is that it provides only "eye-ball" answers to these questions.

—Allison Huck

See also Bar Chart; Line Chart; Pie Chart

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

HISTORIOMETRICS

Historiometrics (or historiometry) is a quantitative method for studying eminent individuals at a distance. The method is most often applied to persons who cannot be subjected to direct psychological assessment. For example, historiometrics was first used to study deceased leaders in a variety of fields. Yet the method may also be applied to persons who are still living but who are otherwise unreceptive to direct psychological examination. For instance, it can be used to study contemporary politicians, such as presidents and prime ministers, who for various reasons may not be willing to volunteer their time (or risk their reputations) to participate in empirical studies. Whether the subjects are alive or not, the method does require that the individuals possess a substantial historical record that can provide a reliable source of data. Thus, historiometrics can be considered a specific form of archival data analysis.

Although the term *historiometrics* was not coined until 1909, the technique itself is among the oldest in the behavioral sciences. The first historiometric investigation was published in 1835 by Adolph Quételet, the same statistician who established the normal curve as the basis for describing individual differences. The first book-length historiometric study was Francis Galton's *Hereditary Genius*, which appeared in 1869. Other pioneering historiometric inquiries were conducted by

James McKeen Cattell, Lewis M. Terman, Edward L. Thorndike, and R. B. Cattell. The majority of the researchers who have applied the technique have been motivated by an interest in individual differences.

Like most other empirical studies, the typical historiometric investigation begins by collecting a representative sample of subjects. However, rather than gathering a random sample from the larger human population, historiometric samples are usually nonrandom. Either the sample exhausts the entire population (e.g., all Nobel laureates, all Olympic gold medalists, all presidents of the United States), or the sample includes the most eminent figures in a given field (e.g., the composers who have had the biggest impact on the classical repertoire, the generals who have fought the most decisive battles). Once the sample is determined, variables can be defined and measured by using raw biographical and other historical information as well as by applying content analysis to speeches, correspondence, or creative products. Frequently, the sampled individuals are measured on the same variables on which more-everyday samples can be evaluated. For instance, historical figures have been assessed on such variables as intelligence or IQ; the NEO Personality Inventory; Cattell's Sixteen Personality Factor Questionnaire; and power, achievement, and affiliation motives derived from the Thematic Apperception Test. Because historiometric data are necessarily correlational, they are then subjected to such statistical techniques as regression analysis, factor analysis, time series analysis, path analysis, and structural equation models.

—Dean Keith Simonton

See also Galton, Sir Francis; Nomothetic Versus Idiographic; Path Analysis; Structural Equation Modeling; Time Series Analysis

Further Reading

Simonton, D. K. (1990). *Psychology, science, and history: An introduction to historiometry*. New Haven, CT: Yale University Press.

Historiometrics articles: <http://psychology.ucdavis.edu/Simonton/dkspdf.html> (especially publications 204 and 257, which provide overviews)

HOLDEN PSYCHOLOGICAL SCREENING INVENTORY

The Holden Psychological Screening Inventory (HPSI, published by Multi-Health Systems, www.mhs.com) is a brief screening instrument designed with contemporary health care demands such as speed and ease of use in mind. It was developed by Ronald R. Holden in response to growing practical concerns involved with other, much larger instruments that require a great deal of time to administer. The HPSI consists of 36 self-report items, each on a 5-point scale; takes approximately 7 minutes to administer; and measures three broad dimensions of psychopathology: psychiatric symptomatology, social symptomatology, and depression. The scales have 12 items each and were developed from factor analysis of 11 of the 12 scales of the Basic Personality Inventory (BPI), which in turn were constructed from the scales of the Differential Personality Inventory, although each test makes use of original items.

The psychiatric symptomatology scale is associated with the BPI scales of hypochondriasis, anxiety, thinking disorder, persecutory ideas, and deviation. This scale can be thought of as measuring general maladjustment and includes items asking about illness, pain, simple coordination (such as standing or walking), tension, and panic. The social symptomatology scale draws from the BPI scales of alienation, interpersonal problems, impulse expression, persecutory ideas, and deviation. It measures externalizing traits and includes items regarding alcohol, drugs, illegal activity, authority, and response to others. The depression scale is related to the BPI scales of depression, social introversion, and self-depreciation. The HPSI depression scale measures general self-esteem and mood. It includes items about life satisfaction, social activity, and feelings of self-worth and ability. The three scale scores are added to yield a total psychopathology score that can then be used as a validity scale, with extremely high or low scores indicating potentially invalid responses. All four of the HPSI scales can be converted to *T* scores (standard

scores with a mean of 50 and standard deviation of 10) or percentiles with reference to normative data.

The HPSI should not be given to children under the age of 14, and caution is recommended with physically disabled or acutely psychotic individuals. Normative data is available for general adult, psychiatric adult, male adult psychiatric offender, high school, and university populations. The test does not require extraordinary skill to administer or score and is amenable to group administration. The HPSI is ideal for identifying people who may require further diagnostic testing or attention as well as monitoring clinical change or progress in the domains measured. It has also been used in the evaluation of treatment programs.

—John R. Reddon and Vincent R. Zalcik

See also Basic Personality Inventory; Personality Tests

Further Reading

Holden, R. R. (1996). *Holden Psychological Screening Inventory (HPSI)*. North Tonawanda, NY: Multi-Health Systems.

Holden, R. R. (2000). Application of the construct heuristic to the screening of psychopathology: The Holden Psychological Screening Inventory (HPSI). In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 97–121). Boston: Kluwer.

Ronald R. Holden's Web site: <http://psyc.queensu.ca/faculty/holden/holden.html>

HOMOGENEITY OF VARIANCE

Homogeneity of variance refers to the assumption that the variance of one variable is stable at all levels of another variable. If a predictor variable is categorical, then the variance of the outcome variable or variables should be the same in each of these categories or groups. With continuous data (such as in cross-sectional research designs), this assumption means that the variance of one variable should be the same for all values of the other variable.

Table 1 Numbers of Things People in Negative or Positive Mood or No Induced Mood Thought They Should Check Before Going on Holiday

	<i>Negative Mood</i>	<i>Positive Mood</i>	<i>No Induced Mood</i>
	7	9	8
	5	12	5
	16	7	11
	13	3	9
	13	10	11
	24	4	10
	20	5	11
	10	4	10
	11	7	7
	7	9	5
$\bar{X}$	12.60	7.00	8.70
S^2	36.27	8.89	5.57

Source: From Davey et al., 2003.

Note: $\bar{X}$ = mean; S^2 = variance.

An Example

One study looked at the processes underlying obsessive-compulsive disorder by inducing a negative mood, a positive mood, or no mood in people and then asking them to imagine they were going on holiday and to generate as many things as they could that they should check before they went away. The data are in Table 1.

Three different groups participated in this experiment, each group representing a level of the independent variable, mood: negative mood, positive mood, and no mood induced. Homogeneity of variance would mean that the variances of scores on the dependent variable (in this case the number of items people generated that needed to be checked) would be approximately the same in the three different mood conditions. For participants in a negative induced mood, the variance of the number of items

generated was 36.27 (see Table 1), but the variances for the positive-mood and no-mood-induced groups were 8.89 and 5.57, respectively.

Testing Homogeneity of Variance

To see whether variances are roughly equal across different levels of a variable, we compare the ratio of the largest and smallest variance. This variance ratio should be less than 2 if homogeneity is to be assumed. In this example, the largest variance is 36.27 and the smallest is 5.57; the variance ratio is $36.27/5.57 = 6.51$, and because this value is greater than 2, homogeneity of variance cannot be assumed.

Homogeneity of variance can also be tested with Levene's test, which tests the null hypothesis that the difference between the variances is zero. If Levene's test is significant at $p \leq .05$, then the variances are significantly different, and homogeneity of variances cannot be assumed. Figure 1 shows the SPSS output for Levene's test for the data in our example; because the significance of the test statistic is less than the conventional level of significance of .05, homogeneity of variance cannot be assumed.

It has been noted that the power of Levene's test to detect differences in variances across levels of a variable depends on the amount of data collected: With large samples, small differences in variances will give rise to a significant Levene's test; conversely, in small samples, relatively large differences between variances may remain undetected.

—Andy P. Field

		<i>Levene Statistic</i>	<i>df1</i>	<i>df2</i>	<i>sig.</i>
Number checks listed	Based on Mean	3.628	2	27	.040
	Based on Median	3.456	2	27	.046
	Based on Median and with adjusted df	3.456	2	15.658	.057
	Based on trimmed mean	3.631	2	27	.040

a. Stop rule condition = As many as can

Figure 1 Test of Homogeneity of Variance^a

See also Dependent Variable; Heteroscedasticity and Homoscedasticity; Independent Variable; Variance

Further Reading

Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). Perseveration of checking thoughts and mood-as-input hypothesis. *Journal of Behavior Therapy & Experimental Psychiatry*, 34, 141–160.

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

HYPERGEOMETRIC DISTRIBUTION

The hypergeometric distribution is used to model the probability of occurrence of events that can be classified into one of two groups (usually defined as success and failure) when sampling from a finite population without replacement. It is a discrete probability distribution, in which the number of possible values taken on by the random variable (the number of successes observed in the sample) is finite, with individual probabilities being between 0 and 1. Further, the sum of the probabilities associated with the random variable's taking on all the possible values is equal to 1. The hypergeometric distribution has applications in acceptance sampling, in which items are produced in finite batches and a decision to accept or reject the batch is made on the basis of a random sample selected from the batch and the observed number of nonconforming items.

Formulas

Suppose a finite population or batch has items that are either conforming or nonconforming. Let us define a nonconforming item to be a “success.” The probability distribution of the number of nonconforming items in the sample, denoted by X , is hypergeometric and is given as follows:

$$P(X = x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}},$$

$$x = 0, 1, 2, \dots, \min(n, D),$$

where

D is the number of nonconforming items in the population,

N is the size of the population,

n is the size of the sample,

x is the number of nonconforming items in the sample, and

$\binom{D}{x}$ is the combination of D items taken x at a time and is given by

$$\frac{D!}{x!(D-x)!},$$

where the factorial of a positive integer x is written as $x!$ and is given by $x(x-1)(x-2) \dots 3.2.1$.

Further, $0!$ is defined to be 1.

The mean or expected value of a hypergeometric random variable is expressed as

$$\mu = E(X) = \frac{nD}{N}.$$

The variance, a measure of dispersion, of a hypergeometric random variable is given by

$$\sigma^2 = \text{Var}(X) = \frac{nD}{N} \left(1 - \frac{D}{N}\right) \left(\frac{N-n}{N-1}\right).$$

Example

A batch of 20 DVDs includes, unknown to the company, 3 nonconforming ones. If an inspector randomly samples 5 DVDs, what is the probability of getting 2 nonconforming DVDs? What are the mean and variance of the number of nonconforming DVDs in the sample?

Using the previously defined notation, and assuming that finding a nonconforming DVD is a “success,” we have $N = 20$, $D = 3$, $n = 5$, and $x = 2$:

$$P(X = 2) = \frac{\binom{3}{2} \binom{17}{3}}{\binom{20}{5}}$$

$$\begin{aligned}
 &= \frac{3!}{2!1!} \cdot \frac{17!}{3!14!} \\
 &= \frac{20!}{5!15!} \\
 &= \frac{(3)(680)}{15504} = 0.1316
 \end{aligned}$$

$$\text{Mean of } X = \mu = \frac{(5)(3)}{20} = 0.75$$

$$\begin{aligned}
 \text{Variance of } X &= \frac{(5)(3)}{20} \left(1 - \frac{3}{20}\right) \left(\frac{20-5}{20-1}\right) \\
 &= 0.5033
 \end{aligned}$$

The complete probability distribution of X , the number of nonconforming DVDs in the sample, may be found using the formula, described previously, for values of X being 0, 1, 2, and 3. The results are shown below:

X	0	1	2	3
$P(X = x)$	0.3991	0.4605	0.1316	0.0088

Note that the sum of all the probabilities equals 1.

Suppose an acceptance sampling plan calls for choosing a random sample of 5 from the batch of 20 DVDs. The batch is to be accepted if no nonconforming DVDs are found in the sample. To determine the chance of accepting the batch, as described previously, we use the calculated probability distribution values. Here, since the batch is accepted if the value of $X = 0$, the acceptance probability of the batch is 0.3991.

Approximations to the Hypergeometric Distribution

When the batch size (N) becomes large, which may also lead to taking larger samples, the combination terms in the hypergeometric distribution probability computation formula may become very large and may cause round-off errors. In such circumstances, an

approximation of the hypergeometric distribution may be considered. When the ratio of the sample size to the population size is small, i.e., $n/N \leq 0.10$ (as a rule of thumb), the binomial distribution serves as a good approximation of the hypergeometric distribution.

—Amitava Mitra

See also Acceptance Sampling

Further Reading

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2006). *Introduction to probability and statistics* (12th ed.). Mason, OH: Brooks/Cole.

Hypergeometric distribution articles: <http://mathworld.wolfram.com/HypergeometricDistribution.html> and http://en.wikipedia.org/wiki/Hypergeometric_distribution

HYPOTHESIS AND HYPOTHESIS TESTING

Suppose a gambler flips a coin and counts 20 heads out of 30 tosses. Unsure whether this is a fair coin with an equal probability of heads or tails or a coin with a bias toward heads, the gambler may reason something like this: If the coin was fair, I would expect 15 heads and 15 tails, but I wouldn't always get this result. Sometimes I might get 16 heads and 14 tails or 13 heads and 17 tails or 20 heads and 10 tails, and so forth. It is even possible that I could get 30 heads and no tails with a fair coin if I was extremely lucky. While I can never be certain, I can make some reasoned statistical arguments about the likelihood of any of the possible combinations of heads and tails. Since I am a gambler, I will gamble in a rational manner. If a given number of heads, say 20 or more, is unlikely enough given my model of random tosses, I will decide the coin is not fair. Otherwise, I will decide that the coin could be fair.

These competing hypotheses about the random or nonrandom nature of obtained statistical results form the basis for *hypothesis testing*. The purpose of hypothesis testing is to make rational decisions about the reality of effects. The basic question is that after collecting

data and describing it using statistical methods, one doesn't know whether the obtained results indicate a real relationship or a chance happening. For example, half the time, by chance, treatment Group A will have a higher mean than control Group B even though the treatment had absolutely no effect. The statistician doesn't want to waste time interpreting results that could have been due to chance (a random generating process). In a like manner, journals want to avoid publishing papers whose results are not real, as science makes little progress when it attempts to give meaning to haphazard or coincidental events. In another sense, policymakers do not want to invest in innovations that do not work, simply because a researcher was unable to distinguish between real and random results. Because chance can never be eliminated as an explanation of a phenomenon, statisticians have developed hypothesis testing procedures to assist in making decisions about the reality of effects. Knowing they can never be right 100% of the time, statisticians have developed procedures to measure the likelihood of many statistical results relative to a chance model and make rational decisions based on that measure.

Hypothesis Testing Procedure

The hypothesis testing procedure can be counterintuitive to many people. It requires constructing a model of what the world would look like given that chance or random processes alone were responsible for the results and that these processes were done an infinite (or at least a very large) number of times. The hypothesis that chance alone is responsible for the results is called the *null hypothesis*, and the model of the result of the application of the random process alone is called the distribution under the null hypothesis. The obtained results are then compared with the theoretical model of the distribution of results under the null hypothesis, and the likelihood of finding the obtained results is measured. This likelihood or probability is called the exact significance level and is the likelihood of finding the obtained result given that the null hypothesis (random model) is true. If the probability of the chance model describing the

obtained results is small, then the results are said to be *statistically significant*. In more technical terminology, the null hypothesis is rejected and the *alternative hypothesis* (that the effects are real, not due to a random process alone) is accepted.

One issue in the hypothesis testing procedure is, How small must small be in order to find statistical significance; for example, should it be .1, .01, or .001? To use the hypothesis testing procedure in a scientifically responsible manner, a criterion, called the *alpha level* or α , must be set before the hypothesis test is performed. A default value of alpha set to $p = .05$ is generally accepted in the social sciences, although other values of alpha should be considered based on the cost of making a decision error. For example, if the cost of deciding the effects are real when they are not is high relative to deciding that the random process could be responsible for the results, then the value of alpha should be set lower than the default value of .05. A common practice is to report both the exact significance level and the level of alpha used to make the statistical decisions but allow the reader to set a different value of alpha to make possibly different statistical decisions if so desired.

The Random Generating Process

The hypothesis testing procedure requires the construction of a model of what the world would look like if a random generating process was solely responsible for the results. There are two commonly accepted methods for construction of such models.

Probability Models

Probability models are a result of a thought experiment done using mathematical techniques. For example, using mathematical methods, it is possible to answer the question "What would the distribution of number of heads look like if I flipped a fair coin 30 times and I repeated this an infinite number of times?" The answer can be found as an application of the binomial theorem and is illustrated in Figure 1.

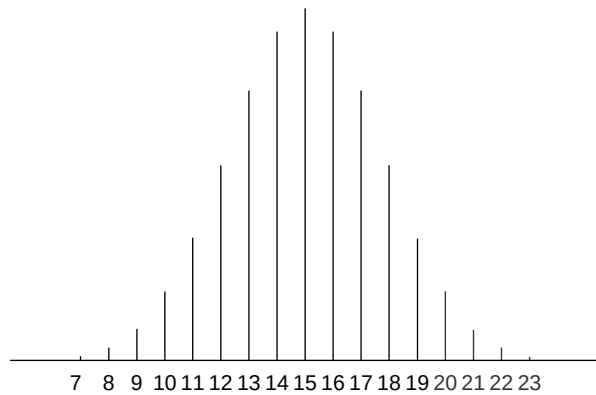


Figure 1 Theoretical Distribution of the Number of Heads in 30 Coin Tosses of a Fair Coin

It is possible using mathematical methods to derive the results of many different types of such thought experiments. *Sampling distributions* or theoretical distributions of sample statistics are extremely useful for such purposes.

Monte Carlo Methods

At times, mathematical thought experiments are not available for hypothesis testing purposes, and brute-force methods must be used. Rather than performing a thought experiment, *Monte Carlo* methods construct probability models with actual or simulated experiments.

It would be possible, but extremely tedious, for example, to take a known fair coin, flip it 30 times, record the number of heads, and repeat the procedure a very large number of times. The distribution of the number of heads could be used as the model of a random process.

A much less tedious method would be to write a computer program that simulated the flipping of 30 fair coins and recorded the number of heads in each set. While not providing the precision of the infinite number of theoretical tosses available in the probability models, this method can often provide a useful approximation of the theoretical models and allows much greater flexibility in the construction of potential models. An example of the application of this

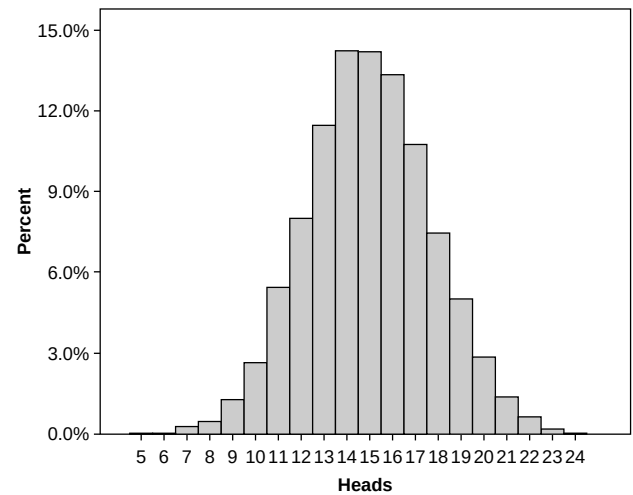


Figure 2 Monte Carlo Distribution of the Number of Heads in 30 Tosses of a Fair Coin

technique is presented in Figure 2 using 10,000 samples of 30 coin tosses.

Decisions

The result of an application of a hypothesis-testing procedure is one of two decisions:

1. It is unlikely the random process generated the obtained results.

In this case, the likelihood of the obtained result given the distribution under the random process (the exact significance level) is smaller than the value that was set for alpha. Alternative ways of expressing this result include the following:

Reject the null hypothesis and accept the alternative hypothesis.

Statistical significance was found.

Real effects were discovered.

2. The results are not unlikely enough to say the random process did not generate the results.

This decision is stated as a double negative because when statisticians discover that the random process could have generated the results, they are not willing to conclude that it did in fact generate the results. This

decision is in fact a type of statistical limbo in which the statistician is not willing to decide either that the effects are real or that they are random. Alternative methods of expressing this decision include the following:

- Retain both the null and the alternative hypothesis.
- Statistical significance was not found.
- No effects were discovered.

A Simple Example Using Hypothesis Testing to Test Whether a Coin Is Fair

A person flips a coin 30 times, obtains 20 heads out of the 30 tosses, and asks “Is this a fair coin, whose likelihood of heads is .50?” Using a hypothesis testing procedure to assist in answering the question, the statistician assumes that the coin is fair (the null hypothesis is true) and proceeds to create a model of what the world would look like in that event (see Figure 1). Before performing the hypothesis test, the statistician sets the alpha level or decision criterion to .05. Comparing the obtained result (20 heads) with what would be expected if chance alone was operating, the statistician concludes that the probability of obtaining 20 or more heads in 30 tosses of a fair coin is .0494, less than the exact significance level. Thus the statistical decision would be to reject the random generating process and conclude that something other than this process was responsible for the results. Note that there is uncertainty about the decision; the statistician knows that 20 heads out of 30 tosses is possible if one is lucky (or unlucky, as the case may be) but that there needs to be a rational basis for making one decision over the other. The hypothesis testing procedure provides that rational basis.

Implications

1. A hypothesis testing procedure can assist in deciding only whether the statistical results are not likely to be due to chance. If statistical significance is found, the hypothesis test generally says nothing

about why chance or randomness was not a good explanation of the results. A careful examination of the data is necessary to explain the results. In some cases, as in a t test, this examination requires a simple comparison of two means. In other cases, such as multiple regression, the pattern of nonrandomness is much more difficult to discern.

2. If statistical significance is not found, the only interpretation is that the selected random process could have generated the results, not that the random process did in fact generate the results. There are any number of reasons that a given obtained result would not be statistically significant, including low statistical power (the probability of discovering real effects when in fact the effects are real) and measurement issues.

3. If a result is statistically significant, or nonrandom, it does not follow that the result has any practical or utilitarian value. In other words, it is possible to discover statistical significance by using a highly sensitive significance test (e.g., one having a very large sample size) and not find the results useful. For example, an expensive new teaching technology might show a statistically significant increase in student performance using a large test sample, but the increase might be too small to justify implementing the technique. For this reason, the results of a hypothesis test should not be presented in isolation but should be reported in conjunction with a measure of the size of the effects.

4. One can never “prove” anything using a hypothesis test. The best one can say is that the selected random generating process is an unlikely explanation of the obtained results, not that the random generating process explanation is impossible. The procedures can provide evidence that a real relationship exists but cannot prove the existence of such a relationship.

—David W. Stockburger

See also Alternative Hypothesis; Null Hypothesis Significance Testing

Further Reading

- Shermer, M. (2002). *Why people believe weird things: Pseudoscience, superstition, and other confusions of our time*. New York: Freeman/Owl Book.
- Stockburger, D. W. (2001). *Introductory statistics: Concepts, models, and applications* (2nd ed.). Cincinnati, OH: Atomic Dog.
- Taleb, N. N. (2001). *Fooled by randomness: The hidden role of chance in life and in the markets*. New York: Texere.
- Web Center for Social Research Methods, by William Trochim: <http://www.socialresearchmethods.net/>

I

During my 18 years I came to bat almost 10,000 times. I struck out about 1,700 times and walked maybe 1,800 times. You figure a ballplayer will average about 500 at bats a season. That means I played seven years without ever hitting the ball.

—Mickey Mantle

ILLINOIS TEST OF PSYCHOLINGUISTIC ABILITIES

The Illinois Test of Psycholinguistic Abilities—Third Edition (ITPA-3) consists of 12 subtests that take approximately 45 to 60 minutes to administer. This edition of the ITPA is designed to assess spoken and written features of language. Norms are provided that permit evaluation of whether a child is developing normally or is at risk for school failure. The scoring system provides information on overall strengths and weaknesses in the general areas of spoken and written language as well as in specific features of language. This information may be used in identifying areas for further in-depth testing or tracking the effects of intervention over time.

The ITPA-3 represents a substantial revision. An experimental version first appeared in 1961, and an expanded version appeared in 1968. Unlike the previous versions, the third edition contains only subtests that measure some aspect of language performance. In addition, the subtests require reading and writing. Thus, the age range has been revised to 5 years,

0 months through 12 years, 11 months of age. Only children 6 years, 6 months of age or older may be administered the written subtests.

Half of the 12 subtests are presented auditorily and require verbal responses. Two subtests assess word knowledge (semantics) by analogies (e.g., birds have nests, lions have ____) and definitions (e.g., I am thinking of something that has paws—possible answer “dog,” “cat”). Two subtests assess knowledge of morphology and grammar through a cloze procedure (e.g., one dog, two ____) and verbatim imitation of grammatical but semantically anomalous sentences (e.g., Cats like to ice skate). Awareness of the phonological (sound) structure of words is assessed in two subtests: one requiring deletion of sounds within a word and another requiring repetition of sequences of rhyming words.

The remaining six subtests require reading and formulation of written responses. Ability to read sentences and comprehend word meaning is required in two subtests: One involves organizing sentences into logical sequences, and a second involves writing nouns that may be combined with specific adjectives (e.g., a giant ____). This latter subtest requires

knowledge of restrictions on the combination of nouns and adjectives. The final four subtests assess the ability to pronounce and spell words with irregular and regular forms.

The scoring system provides composite scores that identify general as well as specific abilities. The raw scores for each subtest may be converted to standard scores and percentile ranks as well as age and grade equivalents. Composite scores for general and specific abilities are derived by adding scores from selected subtests. The composite scores can be converted to quotients and percentile ranks. Both the standard scores for each subtest and the quotients for all of the composite scores may be plotted on a profile that indicates whether the scores are within, above, or below one standard deviation from the mean.

The test is based on Osgood's model of communication, which postulates multiple levels of organization and processing as well as multiple channels of input and output. The test assesses features of linguistic knowledge; expressive and receptive skills, including auditory memory; and some cognitive abilities. Early decoding skills for print-to-sound correspondences and orthographic knowledge are assessed by four of the written subtests.

The test includes new normative data for a sample of 1,522 children that is representative of the 2000 U.S. population of school-aged children. New reliability and validity data are provided. The data indicate that the test has adequate reliability. The extent to which the test predicts school failure, especially when administered to older age groups, still needs to be established. The test is not useful for identifying treatment goals.

—Jennifer R. Hsu

See also Woodcock Johnson Psychoeducational Battery

Further Reading

Osgood, C. E. (1957). A behavioristic analysis of perception and language as cognitive phenomena. In J. S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood, & D. Rapaport (Eds.), *Contemporary approaches to cognition: A report of a symposium at the University of*

Colorado, May 12–14 (pp. 75–118). Cambridge, MA: Harvard University Press.

Towne, R. L. (2003). Review of Illinois Test of Psycholinguistic Abilities—Third Edition. In L. L. Murphy (Ed.), *The fifteenth mental measurements yearbook* (pp. 458–462). Lincoln: University of Nebraska Press.

Psycholinguistics definition: <http://www.answers.com/psycholinguistics&r=67> and <http://en.wikipedia.org/wiki/Psycholinguistics>

IMMEDIATE AND DELAYED MEMORY TASKS

The Immediate and Delayed Memory Tasks (available from Donald M. Dougherty, PhD, at NRLC@wfubmc.edu) are a set of computerized behavioral measures of impulsivity and attention. In the Immediate and Delayed Memory Tasks (IMT/DMT), individuals are instructed to click a mouse button when a target letter or number sequence appears on a computer monitor. Typically, five-digit numbers appear rapidly and briefly. When two consecutive sets of numbers appear and match (e.g., 19701 and 19701), individuals taking the test should click the button. These responses are interpreted as a measure of attention. However, in other trials, the numbers closely match but are not identical. Responses to these trials are interpreted as measures of impulsivity, because they are thought to reflect responding before information processing is completed.

The test is composed of two similar portions, the IMT and DMT, that differ in the length of time required to compare matching numbers. In both tasks, all number sets appear on the computer monitor for 0.5 second, with a 0.5-second break between numbers. In the IMT condition, individuals are instructed to compare successive stimuli, so that they must remember numbers spanning a 0.5-second period. But in the DMT, the numbers to be compared are separated by “distracters,” which are five-digit numbers presented three times between the numbers to be compared. For example, a DMT sequence could read 48593, 12345, 12345, 12345, 48593. As a result,

the IMT requires remembering and comparing data for 0.5 second, whereas in the DMT, it is 3.5 seconds. Typically, the test is administered in four alternating testing blocks of IMT and DMT. The entire sequence takes about 20 minutes to complete.

In addition to the default testing conditions described here, this software package includes adjustable parameters allowing for testing across various samples and conditions. For instance, the IMT/DMT is suitable for assessment of children, adolescents, or adults; also, it is sensitive to state-dependent changes in impulsivity induced by situations such as drug use, medications, or stage of illness. The IMT/DMT is widely used in research settings, and it has excellent psychometric characteristics and extensive validity across a variety of populations.

—Charles W. Mathias, Dawn M. Marsh,
and Donald M. Dougherty

Further Reading

Dougherty, D. M., Marsh, D. M., & Mathias, C. W. (2002). Immediate and Delayed Memory Tasks: A computerized measure of memory, attention, and impulsivity. *Behavioral Research Methods, Instruments, and Computers*, 34, 391–398.

Dougherty, D. M., Mathias, C. W., & Marsh, D. M. (2003). Laboratory measures of impulsivity. In E. F. Coccaro (Ed.), *Aggression: Psychiatric assessment and treatment* (pp. 247–265). New York: Marcel Dekker.

Immediate and Delayed Memory Tasks description and the laboratory where they were developed: <http://www1.wfubmc.edu/psychiatry/Research/NRLC/>

INDEPENDENT VARIABLE

Any research endeavor, accompanied by the use of statistical and measurement tools, often finds one of several different types of variables present, referred to as independent variables.

An independent variable is the variable that is manipulated or changed to gauge the effects on some outcome or dependent variable—it is the “force” that is hypothesized to cause change in the outcome of the

experiment. It is also sometimes called a treatment variable because it is often the treatment or the experimental condition that is applied at different levels (hence, it is a variable) to a selected group, and then the effects of that application are evaluated through an examination of the outcome or the effects on the dependent variable.

For example, one can think of a set of independent variables (in this example, three) acting alone and together (in their interaction) to influence some outcome (usually called an independent variable) as follows:

$$DV = f(IV_1, IV_2, IV_3)$$

where

DV is the value of the dependent variables

f is the function of sign

IV₁ is the first independent variable, with other independent variables (IVs) such as IV₂ and IV₃ to follow.

Table 1 is a simple experimental design where the one independent variable is number of hours of reading per week that a group of children receives in an after-school program. The dependent variable is the children’s comprehension, which is evaluated using a test that assesses that variable.

Each of the three groups (one for each level of the independent variable) receives one of the three treatments, and each level of treatment represents one of the levels of the independent variable, which is labeled Amount of Extra Reading.

Table 1 An Experimental–Control Group Comparison

	Amount of Extra Reading		
	No Extra Reading	5 Hours Extra Reading	10 Hours Extra Reading
Control Group			
Experimental Group			

What are independent variables independent of? One another—the most efficient independent variables are those that act on their own and contribute a better understanding of the dependent variable. Ideally, each independent variable is unrelated to the others, so that its contributions are unique. If the variables are related to one another, it is difficult to separate the effects of each one on the dependent variable and, therefore, to clearly conclude that the change to the dependent variable is a function of any one independent variable rather than a combination.

—Neil J. Salkind

See also Dependent Variable; Moderator Variable

Further Reading

Independent variables further definition and examples: http://en.wikipedia.org/wiki/Independent_variable

Independent and dependent variables further examples and discussion: <http://www.cs.umd.edu/~mstark/exp101/expvars.html>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Mohr, W. K., Lutz, M. N., Fantuzzo, J. W., & Perry, M. A. (2000). Children exposed to family violence: A review of empirical research from a developmental-ecological perspective. *Trauma, Violence, & Abuse*, 1(3), 264–283.

There are many different types of variables, independent or treatment variables being one of them. In this review, Wanda Mohr and her colleagues focused on a developmental-ecological model to evaluate the past decade of research on children exposed to family violence, and a coding system was applied to all empirical studies published in this area since 1987. This coding system allowed for inspection of the current state of the knowledge base from this perspective and the ability to determine the progress that has been made in this research area. A total of 21 quantitative studies and one qualitative study were reviewed. Despite progress over the past 10 years, foundational issues related to definitions of the **independent variable**, substantiation of exposure, developmental sensitivity, and methodology remain. The authors also present recommendations for future research studies.

INDIVIDUALS WITH DISABILITIES EDUCATION ACT

After being enacted in 1975 as the Education for All Handicapped Children Act, the IDEA was reauthorized and revised in 1986, 1990, 1997, and 2004. Moreover, the IDEA is still sometimes referred to as P.L. (Public Law) 94-142, indicating that it was the 142nd piece of legislation introduced during the 94th Congress. The IDEA and its regulations require states, through local educational agencies or school boards, to identify, locate, evaluate, and serve all children with disabilities (34 C.F.R. § 300.125), including those in nonpublic schools, regardless of the severity of their needs. Insofar as the child-find provisions are included as a related service in the IDEA's regulations, many school systems screen preschool children to assist in the early identification of students with disabilities (34 C.F.R. § 451). Changes in the 2004 version of the IDEA now require not only that public school officials identify children who attend nonpublic schools in the districts where they attend classes rather than the districts within which they live, but also that child-find activities for these students in private schools be comparable to those used in public schools. At the same time, public school officials must record and report to state education agencies the number of children from private schools who are evaluated, determined to have disabilities, and served (§ 1412(a)(10)(A)(ii)).

In order to be covered by the IDEA, children must meet three requirements. First, students must be between the ages of 3 and 21 (20 U.S.C.A. § 1412(a)(A)). Second, students must have a specifically identified disability. Third, they must be in need of special education (20 U.S.C.A. § 1401(3)), meaning that the children must be in need of a free appropriate education (20 U.S.C.A. § 1401(8)) in the least restrictive environment that is directed by an individualized education program (IEP) (20 U.S.C.A. § 1401(3)). Related services covers developmental, supportive, or corrective services such as transportation, speech pathology, audiology, psychological services, physical therapy, occupational therapy,

recreation (including therapeutic recreation), social work services, and counseling services (including rehabilitation counseling), among others. Moreover, as children near graduation or begin to age out of special education, school officials must develop individualized transition services plans to promote their movement to postschool activities (20 U.S.C.A. § 1414(d)(1)(A)(vii)(II); 20 U.S.C.A. § 1401(30)).

IEPs must describe students' current levels of educational performance, annual goals, and short-term objectives, the specific services that they will receive, the extent to which they can take part in general education, the date services are to begin and for how long they will be offered, and criteria to evaluate whether they are achieving their goals (20 U.S.C.A. § 1414(d)(1)(A)). IEPs must also discuss how students' disabilities affect their ability to be involved in and progress in inclusive settings and necessary modifications to allow them to take part in the general curriculum. In addition, IEPs must detail any related services that students need to benefit from their IEP (20 U.S.C.A. § 1401(3)(A)(ii)).

The IDEA includes extensive due process protections to protect the rights of qualified children (20 U.S.C.A. § 1415), particularly when dealing with disciplining students with disabilities and determining whether their misbehavior is a manifestation of their disability. Among other protections, parents have the right to take part in developing IEPs that direct the education of their children (20 U.S.C.A. §§ 1414(d)(1)(B)(i) and 1414(f)). In such a case, the Fourth Circuit ruled that parents in South Carolina who chose not to participate in its formulation could not render school officials liable for not having an IEP (*MM ex rel. DM v. School Dist. of Greenville County*, 2002) completed and signed, because their failure to do so was caused by the parents' lack of cooperation. Moreover, officials must provide parents with notice and obtain their consent prior to evaluating or placing children (20 U.S.C.A. § 1414(a)(1)(2), 20 U.S.C.A. § 1414(b)(3)). Once students are placed in special education, school officials must notify parents before trying to change their placements (20 U.S.C.A. § 1415(b)(3)(A)). IEPs must be reviewed at least annually (20 U.S.C.A. § 1414(d)(4)(A)), and children must

be reevaluated completely at least every 3 years (20 U.S.C.A. § 1414(a)(2)(A)). However, a provision in the recently reauthorized IDEA allows up to 15 states to pilot comprehensive, multiyear IEPs that do not exceed 3 years and that are designed to coincide with natural transition points in a child's education (§ 1414(d)(5)(A)(i)). Another change with regard to IEPs permits minor changes to IEPs to be made by means of conference calls or letters (§ 1414(f)). The IDEA also includes provisions, supplemented by the Family Education Rights and Privacy Act (29 U.S.C.A. § 1232g) and its regulations (34 C.F.R. §§ 300.560–577), protecting the confidentiality of all information used in the evaluation, placement, and education of students (20 U.S.C.A. § 1417(c)).

Parents may be entitled to an independent evaluation at public expense if they disagree with a board's evaluation (20 U.S.C.A. § 1415(b)(1)). If parents successfully challenge the assessment of their children, they can be reimbursed for the costs of doing so. Conversely, if a board's evaluation was appropriate, parents are not entitled to further testing at public expense.

At the heart of the IDEA is the requirement that all children with disabilities receive a free, appropriate public education in the least restrictive environment. But because neither the IDEA nor its regulations include a definition of appropriate, it was necessary to seek judicial intervention for such an understanding. In *Board of Education of the Hendrick Hudson Central School District v. Rowley* (hereafter *Rowley*), its first case involving the IDEA, the Supreme Court interpreted "appropriate" as providing a floor of opportunities rather than as a vehicle to maximize a child's potential. In *Rowley*, parents of a kindergarten student in New York who was hearing impaired challenged their school board's refusal to provide their daughter with a sign-language interpreter. A federal trial court and the Second Circuit agreed that officials had to provide the child with an interpreter on the basis that an appropriate education was one that would have allowed her to achieve at about the same level as her peers who were not disabled. The Court, recognizing that the child earned passing marks and advanced academically without the sign-language interpreter, reversed. The Court ruled that

an appropriate education was one that met the IDEA's procedures and was "sufficient to confer some educational benefit (at 200)" on the child. Insofar as the Court was convinced that the child received "some educational benefit" without the sign-language interpreter, the Court concluded that she was not entitled to one even though she might have achieved at a higher level had officials provided her with such assistance.

Rowley's interpreting the IDEA as having set a minimum level of appropriateness does not prevent states from setting higher standards. To date, courts have upheld higher state requirements in such jurisdictions as California, Massachusetts, Michigan, Missouri, New Jersey, and North Carolina. Some of these courts acknowledged that the higher state standards replaced the federal requirements because the IDEA expects special education programs to meet the standards of state educational agencies.

Other courts have interpreted Rowley's "some educational benefit" criterion as requiring more than minimal growth, adding that one must be meaningful or appreciable. Regardless of which criteria apply, under the IDEA's so-called zero-reject approach, reflected most notably by a seminal case from New Hampshire (*Timothy W. v. Rochester, N.H., School Dist.*), school boards must provide services to all eligible children regardless of the severity of their disabilities.

At the same time, students with disabilities must be educated in the least restrictive environment (20 U.S.C.A. § 1401(8)(A)). To this end, the IDEA requires school boards to provide a continuum of alternative placements from least to most restrictive for children with disabilities (34 C.F.R. § 300.551). The first four choices are typically in a child's home school, beginning with the goal of full inclusion in a regular class, to inclusion with help such as a teachers' aide, to partial inclusion, to partial resource room placement, to a self-contained placement in a resource room. The final, more restrictive options are special day schools, hospital or homebound instruction, or residential placements.

In a dispute from California, the Ninth Circuit expanded the existing criteria with regard to least restrictive environment, stipulating that IEP teams had to take four factors into account in placing students:

the educational benefits of placing children in regular classrooms, the nonacademic benefits of such placements, the effect that their presence would have on the teacher and other children in classes, and the costs of inclusionary placements (*Sacramento City Unified School Dist. Bd. of Educ. v. Rachel H.*). Furthermore, in a case dealing with cost, the federal trial court in Utah held that when school officials consider two different programs, either one of which offers an appropriate education, they can take expense into consideration when making a placement (*L.B. and J.B. v. Nebo School Dist.*).

The IDEA's goal of full inclusion notwithstanding, not all students with disabilities must be placed in regular education classes. Courts have approved more restrictive placements where students could not function in regular classes, even with supplementary aids and services, or inclusion did not succeed.

—Charles J. Russo

Further Reading

Board of Education of the Hendrick Hudson Central School District v. Rowley, 458 U.S. 176 (1982).

Family Education Rights and Privacy Act (29 U.S.C.A. § 1232g).

Individuals with Disabilities Education Act (IDEA) (20 U.S.C.A. §§ 1400 *et seq.* *L.B. and J.B. v. Nebo School Dist.*, 214 F. Supp.2d 1172 (D. Utah 2002).

MM ex rel. DM v. School Dist. of Greenville County, 303 F.3d 523 [169 Educ. L. Rep. 59] (4th Cir. 2002).

Osborne, A. G., & Russo, C. J. (2006). *Special education and the law: A guide for practitioners* (2nd ed.). Thousand Oaks, CA: Corwin.

Russo, C. J., Osborne, A. G., & Borreca, E. (2005). Special education update: The 2004 revisions of the IDEA. *School Business Affairs*, 71(5), 41–44.

Sacramento City Unified School Dist. Bd. of Educ. v. Rachel H., 14 F.3d 1398 (9th Cir. 1994).

Timothy W. v. Rochester, N.H., School Dist., 875 F.2d 954 (1st Cir. 1989a), cert. denied 493 U.S. 983 (1989b).

Council for Exceptional Children: <http://www.cec.sped.org>
IDEA and special education updates on regulations, articles, and other general information: <http://www.ed.gov/offices/OSERS/IDEA/>

Office of Special Education Programs information: <http://www.ed.gov/about/offices/list/osers/osep/index.html?src=mr>

INFERENTIAL STATISTICS

Inferential statistics allow researchers to make generalizations about how well results from samples match those for populations. Because samples are parts of populations, samples do not include all of the population information. Thus, no inference can be perfect because samples cannot represent completely their parent populations.

Imagine that a population is defined as all students enrolled in U.S. schools. Suppose researchers want to study how well students in that population enjoy school. A survey is developed, and researchers prepare to collect response data. Researchers estimate the target population consists of 3 million students. Because the population encompasses the entire United States, the team of researchers must include all geographic areas from Maine to Hawaii. However, actually collecting data for all U.S. students is too massive, so a sample is selected. Ideally, this sample would be *randomly* selected, which means that all U.S. students have the same chance of being selected. This makes the sample *representative* of the population. Essentially, this means that although the sample may represent only a small percentage of the students who comprise the complete population, that sample is assumed to reflect the characteristics of all U.S. students, including those not selected.

Parameters, Statistics, and Probability

A key component of inferential statistics is the degree to which error in estimating population values from sample values can be minimized. Probability theory, a branch of mathematics, plays a central role in inferential statistics. Probability theory serves as the backdrop for two important inferential statistics procedures. One is estimation, and the other is hypothesis testing. *Estimation* focuses on the degree to which sample values indicate true population values. For populations, computed values such as means, standard deviations, and variances are called *parameters*. For samples, those values are called *statistics*. Therefore,

questions of estimation address the degree that statistics are equivalent to parameters.

Hypothesis testing pertains to investigators' attempts to answer specific research questions based on theoretical premises. For instance, when researchers want to determine the relationship between two variables—"Is there a relationship between Scholastic Aptitude Test scores and college grade point average?" or "Do males and females differ in their reading ability?"—hypothesis testing converts the research questions into predictive statements so that they can be subjected to empirical testing. Before we describe procedural steps used in hypothesis testing, let's take a very brief look at the history of probability theory in inferential statistics.

Probability Theory and Inferential Statistics

In his classic text titled *Probability, Statistics, and Truth*, Richard von Mises places the onset of probability theory in statistics in the early 1900s. Although properties of distributions, such as those of the normal curve, had been deduced mathematically by the early 1800s, there was limited research on the degree to which the normal curve reflected phenomena observed in the real world. In the early 20th century, the use of probability and the normal curve became important in fields such as agriculture, genetics, and medicine. At this time, R. A. Fisher, a British statistician, introduced the term *likelihood*. This term essentially means probability. Sample data could yield likelihoods of responses that are then compared to what is expected for the population based on properties of mathematical distributions such as the normal curve.

Estimation and Hypothesis Testing Procedures

To illustrate the procedures of estimation and hypothesis testing in inferential statistics, consider the following situation. Suppose high school principals in one district want to answer the following questions about their students' performance on a high-stakes assessment:

1. Did our students perform the same as students nationwide with a mean of 120?
2. Did our students perform differently from students attending a neighboring school district?
3. Did our current students perform better than district students who took the test 4 years ago?

There are four basic inferential statistics steps needed to address any one of the questions. First, researchers must translate each research question into a pair of hypotheses. The first hypothesis, the *null hypothesis*, addresses the question as if the expected answer were “no.” The second hypothesis, the *alternative* or *researcher’s hypothesis*, addresses the question as if the expected answer were “yes.”

In the second step, researchers choose the statistical technique that can help them address each specific question, along with an acceptable error rate with which they justify their conclusion. This error rate is referred to as the *alpha level*. The alpha level (symbolized as α -level) is the degree of *Type I error*. Type I error represents the probability that the null hypothesis will be rejected when the null hypothesis is true for the population. Usually, the alpha level for the social sciences is preset at .05. What this means is that for every 100 times the research question is addressed by a unique sample of data drawn randomly from the same population, there are 5 times when the results are purely attributable to chance.

In the third step of hypothesis testing, researchers actually compute the statistical values given their sample data. Computer programs such as the Statistical Package for the Social Sciences (SPSS) report estimates of the alpha level given the size of the statistical value computed as well as the sample size. Logically, the larger the sample size, the less error in inferring what the population value is based on the sample estimate. This should make sense because larger sample sizes mean that more information about the population is available to the researchers.

In the last step of hypothesis testing, researchers make a decision and state a conclusion. The decision is made in reference to the null hypothesis. Researchers either reject or fail to reject the null hypothesis. If they fail to reject the null hypothesis, it

means the answer to the initial research question was “no.” If they reject the null hypothesis, then they are actually supporting the alternative hypothesis. This means that the answer to the research question is “yes.” The conclusion essentially restates the decision but in less statistical terms. It is usually in their conclusion that researchers also use the word *significant*. This means that if researchers rejected the null hypothesis, then they believe there is a strong likelihood that a result or relationship exists for the population as it does for the sample. We will now take a look at the four steps in answering the principals’ research questions.

To answer Question 1, the researcher would set up the following hypotheses:

$$H_0 : \mu_1 = 120$$

$$H_a : \mu_1 \neq 120$$

Although the notation may be unfamiliar to some readers, the statistical symbols are easy to interpret. The symbol H_0 stands for the null hypothesis, whereas the symbol H_a stands for the alternative hypothesis. For both hypotheses, the Greek letter μ (mu) is used because the high school principals are interested in an average or a mean. All hypotheses must be written to reference parameters, not statistics. When researchers state their conclusion, it is a generalization or inference from the sample results to what is expected for the population. Parameters are symbolized with letters from the Greek alphabet. Therefore, all null and alternative hypotheses should be written in notation form with letters like μ (for means), ρ (rho, for correlation coefficients), or β (beta, for regression coefficients).

In Step 2, the researchers state that a sample mean will be computed for the school district. The alpha level or Type I error level is set at .05. In Step 3, the computed sample mean value is compared to a *critical value* using the mathematical properties of the normal curve. The critical value depends on the sample size and the alpha level, and it is determined based on the theoretical properties of the normal curve. These properties have been derived from mathematical calculations using a formula from calculus. There are infinitely many critical values, just as there are

infinitely many points that represent the score continuum under the normal curve. Mathematicians typically present critical values in appendix tables. Researchers look at their computed sample value and compare it to the critical value given their specific sample size and set alpha level.

In Step 4 of hypothesis testing, if the mean computed for the sample is larger than the critical value, then the researchers reject the null hypothesis. The conclusion would then be drawn that the school district's test score mean is not equal to the nationwide population with its mean of 120.

For Questions 2 and 3, the pairs of hypotheses would be presented in slightly different notational form. For Question 2, two sample means are compared. The first mean represents one school district, and the second mean represents the neighboring school district. Therefore, these two school districts provide two samples of data. We want to know if the two samples represent the same population or different ones. Therefore, the following pair of hypotheses is recorded:

$$\begin{aligned}H_0 : \mu_1 &= \mu_2 \\H_a : \mu_1 &\neq \mu_2\end{aligned}$$

The school principals are still interested in means, so μ s are used to represent population parameters; however, the subscripts 1 and 2 beside the parameters indicate that there are two samples of means. As was done for the first research question, statisticians would proceed with Steps 2 through 4 of the basic set of hypothesis testing procedures. They would state which statistic should be used and then set the alpha level (i.e., Step 2). This time, a t -test value for independent samples would be computed to compare means (i.e., Step 3). The t -test value would be compared to a critical value, and finally, a decision and conclusion about the significance of the results would be made (i.e., Step 4).

It is important to note that for Questions 1 and 2, the investigators used an equality sign when stating the hypotheses. This is because the hypotheses are considered *nondirectional*. This means that if we reject the null hypothesis for the first research question, then we

are inferring only that the mean is different from 120, not whether it is below or above 120. Likewise, for Question 2, if we reject the null hypothesis, then we know only that the two sample means likely represent two different populations. We are not testing which population assumedly has the greater mean.

For Question 3, however, the principals are interested in the direction of the results. Thus, these *directional* hypotheses are written with inequality signs:

$$\begin{aligned}H_0 : \mu_1 &\leq \mu_2 \\H_a : \mu_1 &> \mu_2\end{aligned}$$

where Group 1 consists of the population from which scores from the school are drawn this year, and Group 2 consists of the population from which scores from this school were drawn 4 years ago. If the null hypothesis is rejected, then the researchers will infer that the current average for their school district is significantly greater than the average reported 4 years ago.

Controversies Surrounding Inferential Statistics

There are contemporary debates surrounding the use of inferential statistics. One debate pertains to statistical power. We mentioned that there is always error in inferential statistics, and Type I error is one example. Statisticians must simultaneously deal with *Type II error*, given every null hypothesis tested. Type II error occurs when researchers fail to reject the null hypothesis, and the null hypothesis was wrong. Statistical power is based on Type II error given the simple formula $1 - \beta$, where β stands for the probability of Type II error. Thus, *statistical power* reflects the degree to which a decision is made without error.

Currently, many investigators think that stating the decision and conclusion is insufficient in inferential statistics because statistical power is not addressed directly. What these researchers know is that as sample size increases, there is a greater likelihood of rejecting the null hypothesis. However, the statistical value may be very small and meaningless given researchers' interests. With a large school district, the

test mean might be only 122, and the null hypothesis presented for Question 1 could be rejected. Practically speaking, 122 and 120 are not that different when comparing standardized test scores in schools.

Therefore, many researchers also compute effect sizes. *Effect sizes* tell the direction and magnitude of differences between means in standard deviation units. As stated, as sample size increases, it is easier to reject a null hypothesis. Likewise, in some fields, researchers may not be able to sample large numbers of participants. Consequently, they may not be able to reject the null hypothesis, yet they may be able to compute large effect sizes. For that reason, many journal editors recommend reporting the level of statistical significance *and* the effect size.

—Jonna M. Kulikowich and
Maeghan N. Edwards

See also Hypothesis and Hypothesis Testing; Significance Level

Further Reading

Heiman, G. W. (2006). *Basic statistics for the behavioral sciences*. Boston: Houghton Mifflin.

von Mises, R. (1957). *Probability, statistics, and truth*. New York: Dover.

Inferential statistics: <http://www.socialresearchmethods.net/kb/statinf.htm>

INFORMATION REFERENCED TESTING

A method that has been slow in acceptance by the academic world is an evaluation method called Information Referenced Testing (IRT). Dr. James E. Bruno at the University of California, Los Angeles has been the major proponent of this method. It is not to be confused with Item Response Theory. Item Response Theory is an item analysis method. Although academia has yet to fully embrace this potentially valuable measurement method, industry has been using it for years for purposes of assessment, recertification, and

placement. It has had other names, such as *admissible probability measurement* and *probability scoring*. Currently it is referred to as *confidence-based learning*. Its basis is not new. However, its implementation is made feasible by the advent of high-speed, reliable computers. The computations involved with this method are quite extensive and should be done with a computer. Bruno has provided the measurement community with the tools necessary for using this method. He has written a series of computer programs that will enable a person to carry out the computational aspects of this method, as well as two types of answer forms for recording the responses. One of the forms is a simplification of the other.

A staple within the educational and evaluation process is the multiple-choice test. It is simple to implement and score. In fact, the answer sheets designed for this type of testing work very well with the computer technology in scoring. It is considered objective in that the scoring is preset. There is only one right answer to each question. Contrary to this is the essay test, which is subjectively evaluated but at times has the advantage of allowing the student more freedom to show what he or she really knows. The essay test usually has mixed criteria for evaluation and is costly (time-consuming) to grade. The traditional right-wrong multiple choice is a one-dimensional system. Each student's score is compared either to other students' scores or to some criterion of percent correct. In the traditional multiple-choice-type question, a person with partial knowledge may be undecided between two of the choices, let's say choices *a* and *d* (and say that choice *d* is the correct answer). This person has learned enough to know that the other choices (*b*, *c*, *e*) in the question are definitely not correct. However, the student must make one selection only. If she chooses *a*, she will get the item incorrect. In essence, she would be classified along with the students who picked *b*, *c*, or *e* and suffer the same consequences as they do. Yet she knows more than they do. Wouldn't it be nice if that student could admit that she has only partial knowledge and get to choose both *a* and *d* for a small penalty? Wouldn't it be desirable to have a testing system that allows that and more? By *more*, we mean using a scoring and evaluation system

that could tell each student on which items they have full knowledge, partial knowledge, no knowledge, or the wrong knowledge. IRT, when used properly, can provide these. It could also provide for each student an individualized educational plan that tells them where they are weak and where they are strong. (See Table 1.) Additionally, it would also provide teachers with feedback on what the class as a whole knows and doesn't know. This system can provide both a summative and formative evaluation. When we say wrong knowledge,

we mean that the person has misinformation. Misinformation can be considered more severe than no information because the person usually has a high level of confidence in the wrong answer. Bruno's research and system has provided educators and students with the means of doing this. IRT is a two-dimensional scoring system in which students are measured against an information standard. Tests using this system are objectively scored, and the student's confidence in his or her answer can be assessed.

Table 1 Sample IEP Output

EXAMINEE INDIVIDUAL EDUCATION PLAN (IEP)

EXAMINEE NAME

EXAM NAME

DIAGNOSTIC 90 AUDIT FOR EXPERIMENT

SCHOOL NAME

Cognitive

INSTRUCTOR NAME

PROFESSOR

NUMBER OF QUESTIONS

30

PROCESSING CODE(A<<MCW-APM B-MCW-APM AND RW) = A

FORMATIVE EVALUATION

Examinee misinformation on examination concepts where you were sure of an answer but were wrong. Have instructor explain why the answer you thought was correct was wrong and why another answer was correct.

Item	State	DESCRIPTION—INSTRUCTIONAL CROSS REFERENCE
2	M	Discrimination Between Types of Variables
6	M	Characteristics of Validity: Common Problems
7	M	Confounding of Independent and Dependent Variables
10	M	Identification of Causal Relationships: Balancing
12	M	Use of Random Number Table
18	M	Confounding in an Experiment
24	M	Planned Comparison-Scheffé Test
28	M	Identification of Main Effects From a Table or Graph
29	M	Identification of Interaction Effects From a Table or Graph

Examinee uninformed (lacks information): items on examination concepts where you said you didn't know. Have your instructor explain these concepts to you.

Item	State	DESCRIPTION—INSTRUCTIONAL CROSS REFERENCE
5	U	Validity Definition in Experimental Designs
15	U	Difference Between Randomize vs. Factorial Randomized Group
20	U	Definition of Counter Balance Design
22	U	Recognition of a Reversal Design

(Continued)

(Continued)

Examinee partially informed: Have your instructor review these concepts with you.

Item	State	DESCRIPTION—INSTRUCTIONAL CROSS REFERENCE
4	P	Reliability Definition in Experimental Design
11	P	Definition of a Double Blind Study
16	P	Definition of Quasi-Experimental Design
17	P	Identify Type of Design Used in an Experiment
19	P	Characteristics of an Independent Variable
21	P	Controlling a Single Subject Design
25	P	One-Way ANOVA Definition

Examinee fully informed concepts. Concepts that you said you were sure of the answer and that answer was correct. You have reliable information in these areas. Keep up the good work.

Item	State	DESCRIPTION—INSTRUCTIONAL CROSS REFERENCE
1	I	Definitions Independent and Dependent Variables
3	I	Identification of Independent and Dependent Variables
8	I	Misuse of Correlation Statistic
9	I	Definition of a Control Variable in an Experiment
13	I	Concept of Random Sample and Randomization
14	I	Controlling Extraneous Variables in an Experiment
23	I	Identifying a Type of Research Design
26	I	Definition of Nonparametric Test
27	I	Recognition of Possible and Impossible r Values
30	I	Identification of Simple Effects From a Table or Graph

Student Cognitive Map

Percent	Informed	0.333
Percent	Uninformed	0.133
Percent	Part Informed	0.233
Percent	Misinformed	0.300

With Bruno's development, multiple-choice tests are created with three choices: a , b , and c . However, the test taker is allowed to make one of 13 choices (a , b , c , d , $\dots$, m) with the standard form and seven choices (a , b , c , ab , bc , ac , $?$) on the simplified form. The choice triangle developed by Bruno is given in Figure 1.

The test taker can choose any of the letters. The letters in between a and b allow the test taker to state that he or she feels the right answer is between a and b . Letters closer to a indicate that the person is "leaning" more toward a . The same interpretation applies for the triangle leg a and c and the leg for b and c . Choice m would be chosen if the test taker does not know

enough to answer the question. This choice enables the person to say that he or she "does not know." Unlike other testing methods that have no real correction for guessing, Bruno's IRT does. Test takers are encouraged not to guess and to use choice m if they have to. With choice m , there is no penalty. However, a person who chooses the wrong answer or the wrong leg of the triangle is severely penalized. Using this method of testing, a correct answer is assigned a weight of 30 points, and an incorrect answer is given a weight of -100 points. Bruno and his associates have worked out the theory and mathematics of this; it is complex and beyond the level of discussion here. However, there is proof that

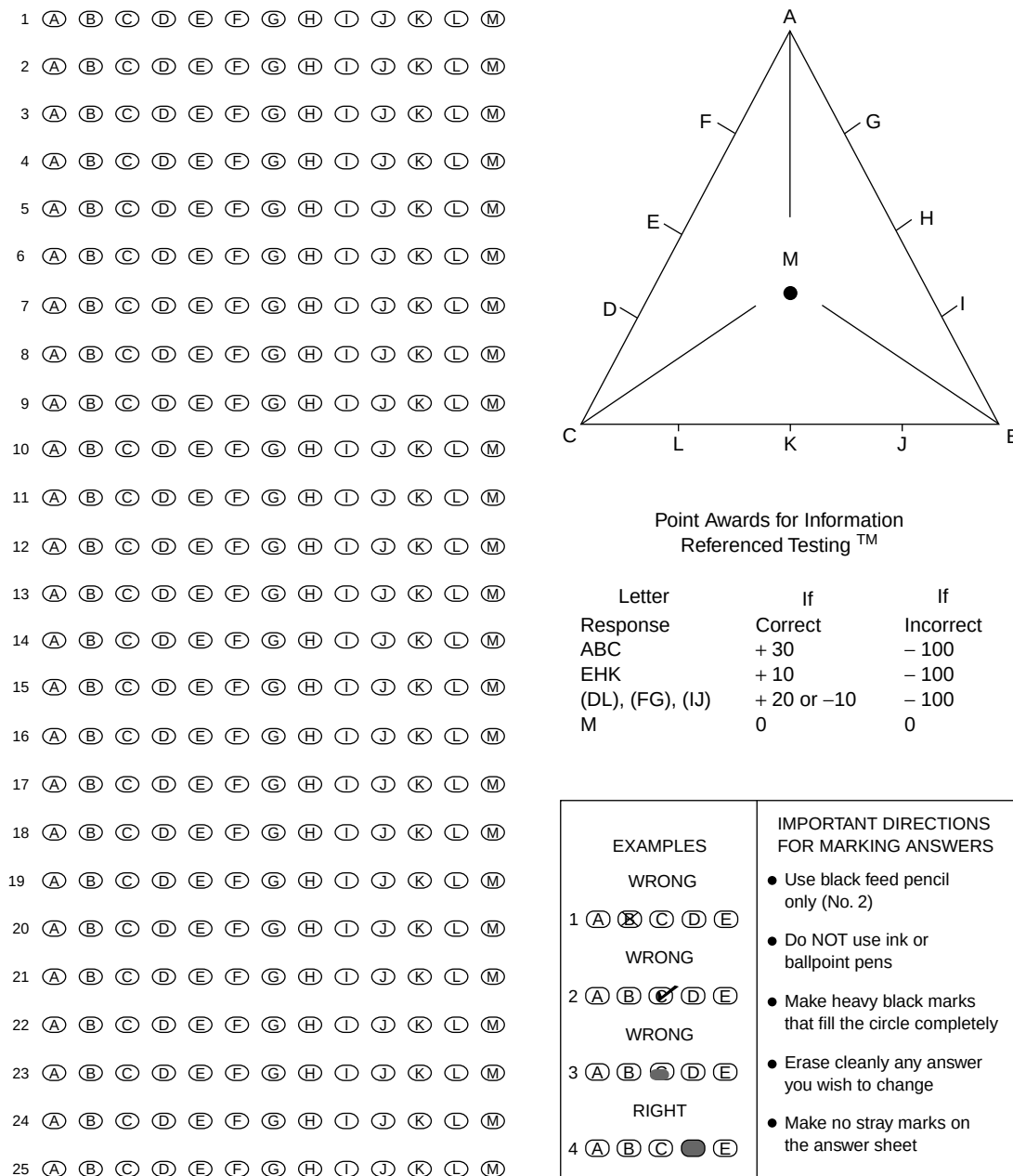


Figure 1 Sample Answer Sheet With Response Triangle

three-choice items are theoretically sound, and that the penalty for obtaining the wrong answer is appropriate. In fact, the mathematics of the scoring system has been worked out so that if the test taker elects not to use the partial information choices and answers every item with an *a*, *b*, or *c* choice, the final score will be exactly equivalent to the traditional right-wrong scoring. Using this method of testing, a correct answer is assigned a weight of 30 points, and an

incorrect answer is given a weight of -100 points. Partially correct answers are assigned +10, +20, or -10 points.

Bruno has successfully applied this method with flight controllers at the Federal Aviation Administration, critical care nursing, and defense systems such as NATO. On the academic level, it has been used to properly place introductory biology students into appropriate labs and discussion sections. It has

helped instructors determine what the students lack in information from prerequisite courses. It has also been used to help the low-achieving student.

The IRT can be considered a controversial method because it is not based on classical test theory or the traditional methods of evaluation and test construction.

—Howard B. Lee

See also Measurement

Further Reading

- Abedi, J., & Bruno, J. E. (1989). Test-retest reliability of computer based MCW-APM test scoring methods. *Journal of Computer-Based Instruction*, 16, 29–35.
- Adams, T. (2005, October). *Eliminating the guesswork in assessment*. Retrieved July 5, 2006, from http://www.wpsmag.com/content/templates/wps_article.asp?articleid=344&zoneid=38
- Bruno, J. E. (1986). Assessing the knowledge base of students with admissible probability measurement (APM): A micro-computer based information theoretic approach to testing. *Measurement and Evaluation in Counseling and Development*, 19, 116–130.
- Bruno, J. E. (1989). Monitoring the academic progress of low achieving students: A analysis of right-wrong (R-W) versus Information Referenced (MCW-APM) formative and summative evaluation procedures. *Journal of Research and Development in Education*, 23, 51–61.
- Bruno, J. E., & Dirkwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational & Psychological Measurement*, 55, 959–966.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Confidence-based learning: <http://www.knowledgefactor.com>
 Information Referenced Testing research program: <http://www.gseis.ucla.edu/faculty/bruno/rassessment.htm>

INFORMATION SYSTEMS INTERACTION READINESS SCALES

The Information Systems Interaction Readiness (ISIR) Scales are a set of self-reported scales that take about 10 minutes to complete and assess user attitudes toward interacting with given information systems (IS). Because the scales provide an overall picture of user experiences related to various mediated actions

in interacting with IS, they can be used to evaluate and compare IS usability in terms of (input) interface, output (interface), and (interaction) rules.

Corresponding to three mediated actions involved in user-system interaction (i.e., using interface to enter input, reading output to receive information, and following rules to communicate with a system), ISIR Scales assess three subconstructs: Input Willingness, Output Receptivity, and Rule Observance. For each, a set of same items measures the affective, cognitive, and behavioral components of user attitudes. To evaluate users' typical feelings, beliefs, and intentions, all items take the semantic differential (SD) form. That is, each item is a pair of bipolar adjectives that elicits responses over a seven-level range in between (from –3 to 3, including zero).

There are 14 different SD items for each subconstruct: six affective, six cognitive, and two behavioral. The selection of affective and cognitive items was based on an understanding of users' typical beliefs and feelings in interacting with IS from a survey study. These items are categorized into Evaluation, Power, and Activity (EPA) dimensions as commonly found in SD scales, in addition to the Intention dimension for the behavioral component (Table 1).

The implications of these dimensions in the context of user-system interaction are specified in the instructions so that subjects can clearly understand the meaning of each item. For example, part of the instructions for Input Willingness scales reads: When I use [system name] interface to enter my input, . . . I **am** _____ to do so (disinclined/inclined; hesitant/ eager); . . . I **feel** _____ toward the interface (dislike/like; rejecting/ accepting); . . . I **find** that the utilization of my input is _____ (foolish/wise; harmful/ beneficial). The wording in the above example indicates, subsequently, the Intention dimension of the behavioral component, the Evaluation dimension of the affective component, and the Power dimension of the cognitive component.

ISIR Scales have exhibited excellent psychometric properties in terms of content, predictive, and construct validities. An overall ISIR score, as well as subscores at different levels, can be calculated by summing the individual item scores or taking simple/weighted averages. The direction (– or +) and magnitude of the ISIR scores indicate whether and how the users are prepared

Table 1 Semantic Differential (SD) Items and Structure for an ISIR Subconstruct

<i>Component</i>	<i>Dimension-Implication</i>	<i>SD Items</i>
Affective	Evaluation—mediator*	dislike/like; rejecting/ accepting
	Activity—operation on mediator**	tense/relaxed; bored/ excited
	Power—goal accomplishment***	annoyed/content; sad/ happy
Cognitive	Evaluation—mediator	useless/useful; imperfect/ perfect
	Activity—operation on mediator	difficult/easy; unsafe/safe
	Power—system cooperativeness	foolish/wise; harmful/ beneficial
Behavioral	Intention—overall mediated action	disinclined/ inclined; hesitant/eager

Notes: *A mediator can be interface, output, or rules. **An operation on mediator can be using interface, reading output, or following rules. ***A goal is the purpose of an operation; for example, using interface is for the purpose of entering input as one wishes. In a similar way, the power dimension for the cognitive component is related to how well users perceive that information systems help them attain the goal.

and willing to interact with given IS. Moreover, the SD items and structure (Table 1) may also be applied to the study of other mediated human activities.

—Jun Sun

See also Semantic Differential Scale

Further Reading

- Crites, S. L., Jr., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, 20(6), 619–634.
- Sun, J. (2005). *User readiness to interact with information systems—A human activity perspective*. In ProQuest Digital Dissertations (www.umi.com).

investigation involving human subjects. In health care settings, informed consent also refers to a communication process used when seeking permission to provide physical or psychological health care services. Informed consent promotes human dignity by upholding the ethical principles of autonomy, beneficence, and justice.

Moreover, effective informed consent procedures enhance public trust in the scientific enterprise. From ethical and legal perspectives, consent is not considered to be truly informed unless the following criteria have been met: (a) disclosure, (b) comprehension, (c) voluntary decision, (d) legal competence, and (e) documentation. This article provides a historical overview followed by a discussion of the five criteria necessary for obtaining informed consent.

Legal protection for human research participants is a recent historical development. In the early 20th century, research and medical consent were typically constrained to verbal assent or a brief written permission agreement. Most physicians and researchers operated under the benevolence model, also known as the “doctor knows best” doctrine. From 1900 to the 1940s, it was thought that lay people were unable to fully comprehend complex procedures. Explanations of intended research objectives and procedures were cursory, at best. Human rights abuses sometimes occurred, and ethnic, cultural, and religious minority populations were particularly vulnerable. The Tuskegee experiment illustrates potential problems when consent is not fully informed. From the 1920s to the early 1960s, a longitudinal study examining the naturalistic course of syphilis was conducted at the Tuskegee Institute. Full disclosure was not provided to the 400 African American male volunteers who were offered free medical treatment in exchange for their participation. Although the participants were aware that the purpose of the study was to examine “bad blood” over several years, they were not explicitly informed that some

INFORMED CONSENT

Informed consent refers to permission granted by a legally competent person to participate in a research

volunteers would be injected with syphilis. Research findings were published in several prestigious medical journals without ethical questions being raised. When eventually confronted, the white investigators argued that they had mistakenly believed that African American recruits understood that “bad blood” specifically referred to syphilis.

During World War II, concentration camp inmates were involuntarily subjected to horrendous medical research, including experimental surgeries, exposure to extreme heat and cold, and other cruel acts. Many painful deaths ensued. These atrocities were later revealed during war crime trials and contributed to the development of the highly influential Nuremberg Code (1946). The Code declared voluntary consent to be a universal human right and also described elements necessary to ensure knowledgeable decision making in what is now known as informed consent. This decisive document had a far-reaching influence on research practice. The Universal Declaration of Human Rights (1948), Declaration of Helsinki (1964), and Belmont Report (1979) further supported human participant rights, including informed consent. Professional ethical codes, such as those developed by the American Psychological Association, also emphasized the need for informed consent. In 1972, government regulation of human research was formalized with the establishment of the Department of Health and Human Services. This agency ensures compliance with federal regulations and policies, including informed consent for human research participation. It also oversees local institutional (or internal) review boards (IRBs) authorized to review research protocols, including informed consent procedures. Currently, informed consent must be obtained before an individual participates in a research protocol unless the project meets federal exemption criteria or has been granted an IRB exception. When written consent is required, the research investigator maintains the responsibility for ensuring that the five fundamental elements of informed consent were met.

Disclosure is the first basic element of the informed consent process. The investigator must be committed to providing the prospective research subject with sufficient information to make a rational

and informed choice about whether or not to participate. The purpose of the research and types of procedures should be described. The classic Milgram Obedience Studies and Zimbardo Stanford Prison Study failed to meet current disclosure expectations; such deception is rarely justified in human research today. If experimental deception is necessary, the investigator should consider alternative procedures. If alternative procedures are untenable, a justification for the reasons why deception is necessary should be prepared for IRB review. Consultation with federal regulations and professional codes is recommended when experimental deception is necessary. After disclosing the aims and procedures, information relevant to the participation risk-benefit analysis is presented. Disclosure of any reasonably foreseeable risks, discomforts, or inconveniences is mandated. If experimental procedures involve more than minimal risk, federal regulations mandate that the investigator provide an explanation of whether compensation or medical treatments will be available if an injury occurs. Then, reasonable benefits of participation are described. The investigator should not amplify or exaggerate the potential benefits. Confidentiality assurances and exceptions, such as intentions to harm self or others, mandated reports of suspected child abuse, or elder/dependent adult abuse, must also be explained. For example, when conducting parent-child communication research, prospective participants should be made that aware that confidentiality would be breached if child abuse is suspected. Court-compelled disclosures should also be explained as a potential confidentiality exception. Research communications on the use of illegal substances could be subpoenaed unless the investigator has obtained special federal confidentiality assurance documentation. Federal regulations also require disclosure of a contact person in the event of questions about the experimental procedures or the participant's rights as a human subject, or to report a research-related injury. Some states impose additional disclosure requirements, such as the name of the sponsor or funding source, the name of an impartial third party for registering complaints, or a separate Participant Bill of Rights form.

Comprehension is the second basic element of the informed consent process. Both oral and written explanations should be easy to understand. Comprehension may be facilitated by presentation of the aims of the research near the beginning of the document, followed by a logical, chronological explanation of the sequence of events. If the procedures are complex or involve substantial time, the use of easy-to-understand and meaningful flow charts, time lines, outlines, or infographics (an illustrational explanation of experimental procedures or process) is highly recommended to enhance understanding. Careful review of the written informed consent document is recommended to ensure that a reading level of 8th grade or below is used for all explanations. Clear definitions for scientific, medical, or legal terminology are needed. If photographs or drawings are included in brochures or infographics, culturally diverse models are recommended. A type size of 12 points or larger will enhance ease in reading for those with visual impairments. During the disclosure process, the prospective participant should be encouraged to read the document thoroughly and ask questions. When the experimental procedure is complex, the potential subject should be encouraged to reread the form. Time to discuss the intended research with family or friends may also be valuable in facilitating understanding and alternative opinions on the intended research. Asking the prospective participant brief questions such as, "Tell me in your own words what this research is doing," "What do you think will happen during this study?" and "What are the risks you might deal with if you take part in this study?" can be useful for determining whether the person truly understands the aims, procedures, risks, and benefits. In some medical research trials, video presentations or interactive computer demonstrations are used to supplement the oral and written explanations of the researchers.

Voluntary decision is the third element in the informed consent process. Coercion, or undue influence, must be avoided at all costs. Although financial reimbursements for time, travel, and other expenses are permissible, the amount should be reasonable. When excessive, such reimbursements will be

viewed as undermining autonomous decision making. For incarcerated populations, research participation should not be tied to favorable parole judgments. Federal regulations mandate that the written statement must explicitly state that participation is voluntary and that refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled. A parallel statement that subjects may discontinue participation at any time without penalty or loss of benefits is also required by federal regulations. Once the information has been disclosed and comprehended, the prospective participant can decide whether or not to participate. Ample time should be provided so that the prospective subject can fully evaluate the information presented and engage in a meaningful risk-benefit utility analysis. Some ethicists advocate that there be at least a one-day deferral between the time the study is explained and the actual granting of written consent by the prospective participant. Investigators uphold the ethical principle of autonomy by fully supporting the invited party's independent decision making and choice to decline or participate in the research protocol.

Competence to make an informed decision is the fourth element of the consent process. It refers to the ability to understand the aims and procedures as well as the ability to make a rational utility judgment that weighs anticipated risks relative to anticipated benefits. Legal capacity is determined by a court of law, rather than investigator opinion or judgment. Adults, aged 18 and above, are typically capable of making financial and health care decisions unless cognitively impaired. A legal or personal representative may be assigned when a person has long-lasting cognitive impairments (such as dementia, severe brain damage, or moderate to profound mental retardation) or decision making is temporarily impaired (active suicide ideation, imminent violence potential, or currently psychotic). In such cases, the legal jurisdiction may permit substitute consent by the party who is legally responsible for the impaired person. Moreover, minors, unless legally emancipated, require the signature of a parent or legal guardian on informed consent documents. When working with minors, it is

important to consult state regulations because minors are sometimes permitted to consent to medical or psychological outpatient care under certain limited circumstances. In cases of substitute consent, it is recommended that the investigator attempt to disclose information about the study and, when possible, follow the desires and preferences of the legally incapacitated party.

Consent, or approval to participate, is the fifth essential element of the informed consent process. In most cases, IRB approval will require written documentation that permission was granted. However, an IRB may waive written documentation in accordance with certain government exceptions (see 45 CFR 46.17 and 21 CFR 56.109). When written documentation is required, approval by the subject or legally authorized representative is indicated by a signature and date. The investigator retains the original document while providing a copy to the subject or legally authorized representative.

Some bioethicists argue that the informed consent process is inherently flawed. In complex clinical trials often found for new medical procedures or consideration of investigational medical devices, the invited subject may not fully understand the medical information presented. Moreover, consent forms may run four pages or longer. In such cases, the prospective participant may not fully engage in a personal risk-benefit utility analysis. The prospective participant may rely on a credible investigator or a trusted physician. This may lead to a decision to agree to participate in the research protocol without fully understanding or weighing out the consequences. Furthermore, clinical research trials often are held in large university-affiliated hospitals. Prospective subjects may trust the integrity of the university credentials and feel an altruistic expectation to participate. Furthermore, a lengthy and complex research procedure requires patience and time to evaluate. In some settings, there may be pressure to shorten the decision-making time if a frustrated person says, "I'll sign these papers even though I don't understand exactly what you'll be doing." Despite these potential problems, disclosure, comprehension, voluntary decision, legal competence, and documented permission

are invaluable in promoting human dignity and maintaining confidence in the integrity of scientific endeavors.

—Carolyn Brodbeck

See also Ethical Issues in Testing; Ethical Principles in the Conduct of Research With Human Participants

Further Reading

- American Medical Association. (2004). *Code of medical ethics: Current opinions with annotations, 2004–2005*. Chicago: Author.
- American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.
- Arnold, R., & Lidz, C. (1995). Informed consent: Clinical aspects of consent in health care. In W. Reich (Ed.), *Encyclopedia of bioethics* (Vol. 3, Rev. ed., pp. 1250–1256). New York: Macmillan.
- Beauchamp, T., & Childress, J. (1994). *Principles of biomedical ethics* (4th ed.). New York: Oxford University Press.
- Faden, R., & Beauchamp, T. (1986). *A history and theory of informed consent*. New York: Oxford University Press.
- Fischman, M. (2000). Informed consent. In B. Sales & F. Susan (Eds.), *Ethics in research with human participants* (pp. 35–48). Washington, DC: American Psychological Association.
- U.S. Food and Drug Administration. (1998). *Information sheets: A guide to informed consent*. Retrieved on August 27, 2005, from <http://www.fda.gov/oc/ohrt/irbs/informed-consent.html>
- Simplification of informed consent documents: <http://www.cancer.gov/clinicaltrials/understanding/simplification-of-informed-consent-docs>

INSTRUMENTAL VARIABLES

For a linear regression model, the consistency of the ordinary least squares (OLS) estimator depends heavily on the assumption that the explanatory variables and the statistical disturbances are uncorrelated. When the regressors are uncorrelated with the disturbance term, we say that we have *exogenous* explanatory variables, whereas a regressor correlated with the error term is said to be *endogenous*. The terms

exogenous and *endogenous* originated in simultaneous equations analysis, but the term *endogenous explanatory variable* covers any case where a regressor is correlated with the disturbance term. A usual source of endogeneity is omission of important variables. Other sources include simultaneity with one equation forgotten and autoregressive models with serially correlated errors.

When an explanatory variable is endogenous, it is not plausible to separate variation in the explanatory variable from variation in the disturbance term, and as a result, OLS yields a biased and inconsistent estimator. In order to separate the variation in the explanatory variables from the variation in the error term, we need more information, called *instrumental variables* or simply *instruments*. In linear models, any variable uncorrelated with the error term is an instrumental variable.

Example

Suppose that we are interested in estimation of the demand curve for a good, and gather data for the price P and the quantity purchased Q of the good. One might set up the following model for the demand curve:

$$\ln Q^D = \alpha + \beta \ln P + u^D,$$

and run a regression of log quantity on log price. However, although the above equation describes the functional relationship between the price P and the quantity demanded Q^D , the collected price and quantity data are *equilibrium* prices and *equilibrium* quantities, that is, the solutions to the simultaneous equations

$$\ln Q^D = \alpha + \beta \ln P + u^D,$$

$$\ln Q^S = \gamma + \delta \ln P + u^S,$$

$$\ln Q^D = \ln Q^S,$$

where the first equation is the demand equation, and the second is for supply. In other words, although we are interested in the slope of the demand curves D_1 , D_2 , D_3 , and so on in Figure 1, the observed data are the equilibria E_1 , E_2 , E_3 , and so on, which are affected by the demand shocks u^D and the supply shocks u^S . Therefore, least squares regression using the observed

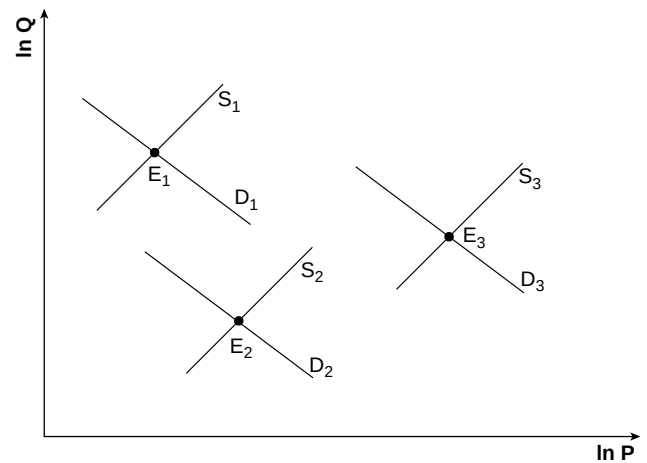


Figure 1 Demand and Supply When Both Demand and Supply Shocks Are Present

data is unlikely to yield an unbiased estimator for the demand function.

An intuitive solution to this problem is to consider a third variable that shifts the supply curve but does not affect the demand. Then, the equilibrium prices and quantities trace out the demand curve as shown in Figure 2, and proper exploitation of the additional variable will lead to a consistent estimator. This variable, the instrumental variable, is correlated with the price P , but is uncorrelated with the demand shocks u^D .

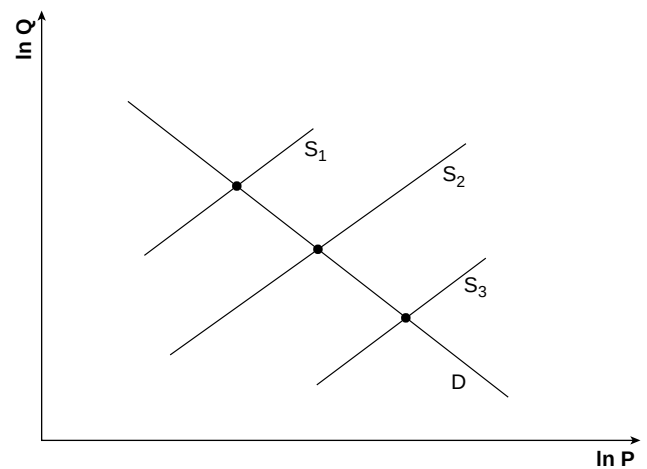


Figure 2 Demand and Supply With Only Supply Shocks

Note: A variable that affects only supply is an instrumental variable.

Instrumental Variable Estimation

Consider the linear regression model

$$y_t = x_t' \beta + u_t = x_{1t}' \beta_1 + x_{2t}' \beta_2 + u_t, \quad t = 1, \dots, T,$$

where the variables in x_{1t} are exogenous and x_{2t} is the vector of endogenous regressors. By definition, the exogenous variables x_{1t} are all uncorrelated with the error term, so these are all instruments. In addition, we may consider collecting additional variables z_{2t} , which are also uncorrelated with the errors. Together, the instruments are $z_t = (x_{1t}', z_{2t}')$, and by assumption, these are uncorrelated with the errors u_t ; that is, $Ez_t u_t = 0$, or equivalently,

$$Ez_t(y_t - x_t' \beta) = 0. \quad (1)$$

We assume that this condition is satisfied at a unique parameter vector β , in which case β is said to be *identified* by the instruments. We also assume that the variance-covariance matrix of z_t is nonsingular, meaning that no elements of z_t are perfectly correlated.

Exactly Identified Case

If the number of instruments is the same as the total number of explanatory variables (exogenous and endogenous), then the system is said to be *exactly identified*, and we can estimate β using the method of moments based on Equation 1:

$$\frac{1}{T} \sum_{t=1}^T z_t(y_t - x_t' \hat{\beta}) = 0.$$

An explicit form solution is available in this case, with

$$\hat{\beta} = \left(\sum_{t=1}^T z_t x_t' \right)^{-1} \sum_{t=1}^T z_t y_t,$$

where $\sum_{t=1}^T z_t x_t'$ is required to be invertible. This estimator $\hat{\beta}$ is called the *instrumental variable (IV) estimator*.

Two-Stage Least Squares

When there are more instruments than explanatory variables, the system is said to be *overidentified*, and there are a few distinct ways of generalizing IV estimation. The most widely used method is *two-stage least squares* (TSLS) estimation or *generalized instrumental variable* (GIV) estimation, which estimates the parameter β using the following two-stage procedure:

1. First-stage regression: Regress each endogenous explanatory variable individually on the instruments (i.e., the exogenous regressors and any additional instruments), and compute its fitted values.
2. Second-stage regression: Regress the dependent variable y_t on the exogenous regressors and the fitted values for each of the endogenous variables obtained from the first-stage regressions; that is, using OLS, estimate

$$y_t = x_{1t}' \beta_1 + \hat{x}_{2t}' \beta_2 + u_t,$$

where $\hat{x}_{2t}$ are the fitted values from the first-stage regressions. The TSLS estimator is the estimator $\hat{\beta}' = (\hat{\beta}_1', \hat{\beta}_2')$ from the second-stage regression.

Letting X , y , and Z denote the matrix of regressors, the vector of the dependent variable, and the matrix of instruments, respectively, the TSLS estimator is

$$\hat{\beta} = (X' P_Z X)^{-1} X' P_Z y, \quad (2)$$

where $P_Z = Z(Z'Z)^{-1}Z'$. When the number of instruments equals the number of exogenous and endogenous variables (i.e., the number of additional instruments equals the number of endogenous variables), the TSLS estimator is algebraically equivalent to the IV estimator of the exactly identified case.

Limited Information Maximum Likelihood

The TSLS estimator makes use of the information that the instruments are uncorrelated with the disturbance term. We can make a further assumption that the endogenous variables are jointly normally

distributed and use maximum likelihood estimation under proper constraints. The resulting estimator is a *limited information maximum likelihood* (LIML) estimator, which was derived by Anderson and Rubin in 1949 in their paper published in the *Annals of Mathematical Statistics*. In matrix form, the model is

$$y = X_1\beta_1 + X_2\beta_2 + u,$$

where X_1 is the matrix of exogenous regressors and X_2 is the matrix of endogenous regressors (all with T rows). Let Z_2 be the matrix (with T rows) of extra instrumental variables. Then the LIML estimator for β_2 equals the *least variance ratio* estimator, which minimizes the variance ratio

$$\ell = \frac{(y - X_2\beta_2)'M_{X_1}(y - X_2\beta_2)}{(y - X_2\beta_2)'M_{[X_1, Z_2]}(y - X_2\beta_2)},$$

where $M_A = 1 - A(A'A)^{-1}A'$ for any matrix A such that $(A'A)^{-1}$ exists. The LIML estimator for β_1 is $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2)$, where $\hat{\beta}_2$ is the LIML estimator for β_2 .

When the number of instruments is equal to the number of regressors, the LIML estimator is identical to both the IV (of the exactly identified case) and TSLS estimators. However, when there are more instruments than regressors, the LIML estimator is different from the TSLS estimator. The TSLS and LIML estimators have the same limit distribution, and therefore TSLS is efficient when the normality assumption is true. Because of the efficiency of TSLS, its simplicity, and because it relies on much weaker assumptions, TSLS is usually preferred by practitioners. However, when the correlation between the endogenous regressors and the extra instruments is weak, it has been shown that LIML is less biased than TSLS, so in this case, the LIML estimator is preferred.

In the case where the estimation equation is nonlinear in the parameters—for example, $u_t = f(x_t, y_t, \beta)$ with $Ez_t u_t = 0$ —we use *nonlinear* IV estimation, which is an application of generalized method of moments based on the moment conditions

$Ez_t f(x_t, y_t, \beta) = 0$. When a system of equations is considered, the TSLS estimation is extended to *three-stage least squares* estimation, and the LIML estimation to *full information maximum likelihood* (FIML) estimation. Details can be found in William Greene's econometrics textbook, among others.

Properties of the IV Estimators

The sampling distribution of the TSLS estimator is straightforwardly obtained from Equation 2. More specifically, under the condition that the observations are *iid* and $Ez_t u_t^2 z_t' = \sigma^2 \Sigma_{zz}$, where $\sigma^2 = Eu_t^2$ and $\Sigma_{zz} = Ez_t z_t'$, we have $\hat{\beta} \rightarrow_p \beta$ and

$$\sqrt{T}(\hat{\beta} - \beta) \sim N(0, \sigma^2(\Sigma_{zx}' \Sigma_{zz}^{-1} \Sigma_{zx})^{-1}),$$

where σ^2 is consistently estimated by $\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (y_t - x_t' \hat{\beta})^2$, and Σ_{zz} and $\Sigma_{zx} = Ez_t x_t'$ are consistently estimated by $T^{-1} \sum_{t=1}^T z_t z_t'$ and $T^{-1} \sum_{t=1}^T z_t x_t'$, respectively. Asymptotically valid tests are implemented naturally from these results. The asymptotic distribution of the LIML estimator is the same as that of the TSLS estimator.

For the above asymptotic behavior of the IV estimators (such as consistency and asymptotic normality), we require that the limit Σ_{zx} of $T^{-1} \sum_{t=1}^T z_t x_t'$ be nonsingular. If the observations are *iid*, then it follows that $Ez_t x_t'$ is nonsingular, that is, each instrument has some nonnegligible explanatory power for at least one regressor (with the other instruments held fixed), and all the regressors are sufficiently explained by the instruments.

When the additional instruments z_{2t} are weakly correlated with the endogenous regressors, it follows that Σ_{zx} is nearly singular, and the problem of so-called weak instruments arises. In this case, the IV estimators are biased and the inferences based on the above asymptotic results are invalid. In addition to the presence of weak instruments, if some instruments are mildly endogenous (i.e., they are slightly correlated with the errors), the bias of the IV estimator may be larger than the bias of the OLS estimator, and therefore special caution is required. The problem of weak instruments is well explained in J. Bound, D. Jaeger,

and R. Baker's paper in the *Journal of the American Statistical Association*.

Specification Tests

Testing the Validity of Instruments

The consistency and limit normality properties of the IV estimators rely on the validity of the instruments. In particular, we require that (a) the instruments are uncorrelated with the errors (legitimacy) and (b) the instruments and the regressors are strongly correlated (relevance).

When there is a single endogenous regressor (together with an arbitrary number of exogenous regressors), relevance of the instruments reduces to the condition that the extra instruments and the endogenous regressor are strongly correlated after the exogenous regressors are controlled for. This can be tested statistically by regressing the endogenous regressor on all the instruments and testing the null hypothesis that the extra instruments z_{2t} have no impact on the endogenous regressor. More specifically, let x_{2t} be the single endogenous regressor and $z_t = (x'_{1t}, z'_{2t})'$ be the instruments. The first-stage regression is run on

$$x_{2t} = \pi'_1 x_{1t} + \pi'_2 z_{2t} + v_t,$$

and an F test is conducted to test $H_0: \pi_2 = 0$. If this null hypothesis is rejected with a large F test statistic, then the extra instruments are relevant for the endogenous regressor. This test is usually called the *first stage F test*.

Testing the legitimacy of the instruments ($H_0: Ez_t u_t = 0$) is also possible if there are more instruments than regressors, that is, if the system is *overidentified*. A widely used method is an LM test, which computes TR^2 , where T is the sample size and R^2 is the R -squared in the regression of the residuals $\hat{u}_t = y_t - x'_t \hat{\beta}$ on the exogenous variables z_t . Here, $\hat{\beta}$ may be the TSLS estimator, or any other efficient estimator, such as LIML. The resulting LM test statistic is approximately χ^2 distributed with degrees of freedom equal to the degree of overidentification, that is, the number of instruments minus the number of regressors. A large value is taken as evidence that some instruments

are not exogenous. This test is available only under overidentification, which is the reason the test is called the *overidentification test*. Under exact identification, the test statistic is equal to zero, so no testing is available.

Testing Endogeneity of Regressors

If all the regressors are exogenous, then the OLS estimator is consistent (and so are the instrumental variable estimators). If some regressors are endogenous, then the OLS estimator is inconsistent while the IV estimators remain consistent. However, because the instrumental variable estimators make use of information about the regressors indirectly through the instruments, whereas the OLS estimator directly uses the regressors themselves, when there is no endogeneity, the OLS estimator is more efficient than the IV estimators. Therefore, it is useful to test whether or not the regressors are exogenous ($H_0: Ex_t u_t = 0$).

A general test from Hausman that compares the OLS estimator and the TSLS estimator is available, because

$$(\hat{\beta}_{TSLS} - \hat{\beta}_{OLS})' (\hat{V}_{TSLS} - \hat{V}_{OLS})^{-1} (\hat{\beta}_{TSLS} - \hat{\beta}_{OLS}) \rightarrow \chi_K^2,$$

where $\hat{V}_{OLS}$ and $\hat{V}_{TSLS}$ are the estimated covariance matrices of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{TSLS}$, respectively, and K is the number of regressors. If the test statistic is large, we conclude that some regressors are endogenous and that the OLS estimator is inconsistent; if the test statistic is small, there is no endogeneity and OLS is consistent.

When only one regressor is endogenous, a simplified test is available based on the idea that the part of the endogenous regressor correlated with the error term is in the disturbance term of the first-stage regression equation. This test is implemented by performing a simple t test for $H_0: \delta = 0$ for the regression

$$y_t = x'_t \beta + \delta \hat{v}_t + \text{error}_t,$$

where $\hat{v}_t$ are the first-stage regression residuals.

—Chirok Han and John Randal

See also Generalized Method of Moments

Further Reading

- Anderson, T., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20, 46–63.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variable estimation when the correlation between the instruments and endogenous explanatory variables is weak. *Journal of the American Statistical Association*, 90, 443–450.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4), 765–799.
- Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics*. Reading, MA: Addison-Wesley.
- Wooldridge, J. M. (2005). *Introductory econometrics* (3rd ed.). Belmont, CA: Thomson.

INTELLIGENCE QUOTIENT

Few terms generate as much anxiety, or as much controversy, as *intelligence quotient* (IQ). The term is emotionally loaded and carries with it the connotation of a fixed entity that resides in each person to varying degrees and that, to a large extent, determines a person's worth in the larger society. Yet, as so often happens, the connotation of IQ today, 100 years after the term was coined, represents something quite different from the original meaning and spirit behind the phrase.

Historical Background

Before Alfred Binet (1857–1911) and, to some extent, Francis Galton (1822–1911), gradations of mental worth were generally determined by physiological indices, such as cranial capacity. Galton extended the psychometric assessment of human qualities in many directions, including the assessment of intellect with various response time measures; however, Binet's work had the single most significant impact on the conceptualization of IQ.

Binet was commissioned by the French government to assist with the identification of students who were unlikely to benefit from ordinary schooling and

therefore should be offered remedial or special education. Having become quite discontented with the utility of cranial measures, Binet searched for something more definitive. His early ideas for developing a test of intelligence drew heavily from one of his countrymen, Binet, who had developed a series of structured questions that were designed to assess the judgment abilities of the individual.

It is important to note that Binet was very explicit in stating that the scores derived from his tests were rough, that they were not intended for use in ranking normal children, and above all else, they were indicators of current functioning and did not speak to the past or future capabilities of the child. An educator at heart, Binet was a strong believer in cognitive modifiability, a view that suggests that intelligence is not a fixed quantity, but one that can be modified and enhanced. This view tends to sit in stark contradiction with many modern theories of intellectual ability, which suggest that intelligence is an innate and relatively fixed capacity.

As part of his remedial education programs, Binet advocated what he called exercises in “mental orthopaedics.” These were based on the belief that one first needed to learn how to learn. He linked increased academic performance as a function of training to an increase in intelligence. Binet was also concerned that scores on his tests should not be misinterpreted, and he cautioned overzealous teachers against the temptation to use the test results to get rid of unruly or uninterested students.

The Problems of Measuring Intelligence: The Appearance of Mental Age

Binet recognized that when he added up the marks on his scales, the score in and of itself was unable to tell him very much about the ability of the individual. What was needed was some way to compare a child's score with some benchmark. Binet was particularly interested in disentangling native intelligence from the effects of schooling; thus, tests of educational achievement would not serve as an appropriate comparison. Binet recognized that this benchmark needed to be empirical, because he was rightfully cautious

about accepting the evaluations of parents and teachers too literally, believing they were susceptible to any number of biases (e.g., the protective parent might exaggerate the capacity of his or her child; a teacher wanting to minimize troubles in his or her classroom may provide an underwhelming evaluation of the child). Furthermore, without some clearly defensible and replicable criterion that could be agreed upon, meaningful assessment of change would not be possible.

With this in mind, Binet set out to establish a replicable, empirical criterion for grading intelligence. The first problem that needed to be addressed was the generally accepted observation that intelligent *behavior* tends to increase with age. The line of reasoning that Binet pursued to deal with this is very illuminating and, although subsequently modified and improved, serves as an exemplary model of the application of sound logic and experimentation to the investigation of human ability. Binet, along with his student Theodore Simon, administered his tests to children of different ages and collated their scores. They then ranked the tests according to difficulty. Out of this work, the term *mental age* was coined. Mental age was defined as the average age of a child of normal intelligence who could pass the test. It is significant that the chosen comparisons were normally functioning children. Hence, regardless of his or her age, a child who was able to pass a test that a normal child was able to pass at 5 years old was assigned a mental age of 5. This meticulous process of establishing benchmarks for comparing individuals still informs modern methods of test development, albeit in more sophisticated forms. We call this process *test standardization*, and the benchmarks *norms*.

Problems With Mental Age and the Advent of the IQ

Almost immediately, mental age ran into problems. Consider two individuals with the same mental age of 9 years; the first is 18 years old and the second is 6. The 18-year-old is likely to think qualitatively differently from the 6-year-old, even though they have the same mental age—their behaviors, judgments, and

processes are simply different. The solution as to how to evaluate such differences meaningfully was to consider the ratio of mental age to chronological age—this ratio was called the intelligence quotient, or IQ. Using this transformation, an 18-year-old with a mental age of 9 would have an intelligence quotient of 0.50. The 6-year-old with the same mental age would have an intelligence quotient of 1.50. To remove the decimal and simplify reporting, the quotient was multiplied by 100, and hence, the 18-year-old was assigned an IQ of 50, and the 6-year-old an IQ of 150. The lay person's understanding of the IQ scale has generally stayed the same ever since.

A more fundamental problem influenced the reliable application of IQ scores based on mental age. This problem was particularly apparent when, contrary to Binet's original intentions, the scale was modified for use in the United States to rank normal and superior children and adults. As mentioned earlier, performance on the cognitive tests of Binet increased with age. But intelligence does not keep increasing indefinitely. At some point, an increase in age will not contribute to any significant improvement in performance on the test. Hence, to apply the IQ to adults, the chronological age needed to be truncated at a suitable point to allow the IQ scale to remain meaningful. Determining the appropriate point to truncate was complicated by the fact that the mental age of a test is determined by its difficulty. An easy test will have a low mental age and a harder test a higher mental age. Researchers presented various arguments to support their chosen point of truncation, and for practical purposes, this seemed to address the problem. However, the IQ score has an even more troublesome limitation.

Problems With IQ and the Advent of the Deviation IQ

One of the proposed benefits of the IQ over mental age was that it facilitated comparison of individuals at different ages. It was recognized that the useful interpretation of IQ required a constancy of scores across ages. That is, the ratio of mental age to chronological age should be the same (i.e., 100) for a normal functioning child no matter how old he or she is. However,

Wechsler reports that in the 1937 version of the Stanford-Binet intelligence scale, there was considerable variability in mean IQ scores across different ages—the mean IQ for 2½-year-old children was 109.9, but only 100.5 at 14 years. Furthermore, Wechsler argued that the variability (standard deviation) in IQ scores also differed considerably at different ages. The standard deviation in IQ scores for 12-year-olds was 20 IQ points, yet the standard deviation was only 12.5 points for 6-year-olds. The reason for this variation might be interpreted to lie with the particular characteristics of the standardization sample chosen as the comparison group, and there is evidence that this was recognized by researchers at the time. However, the problem this presented remained a significant obstacle to the reliable interpretation of IQ scores and had the potential to render any comparisons of individuals over time or at different ages meaningless.

The solution was an act of statistical ingenuity. As Wechsler describes, scores on the test were just that—scores. The numbers used to represent the scale are always arbitrary in the sense that there is no fixed point of origin—zero on the scale did not imply no intelligence, and each test's mean score depended on its difficulty. Hence, mental age and IQ were just numbers on some arbitrary scale determined by the researcher. However, it is possible to mathematically transform the scores on one scale to any other scale without changing the rank ordering of individuals or the *relative* distance between them. This is what we do when we convert raw scores to standardized scores (i.e., *z* scores) in, for instance, the process of testing for statistically significant differences between two population means. *z* scores always have a mean of zero and a standard deviation of one, regardless of what the original raw score scale is. Furthermore, if we have two groups of people, say, 5- and 10-year-olds, and we convert the raw scores of each group separately to *z* scores, then the mean and standard deviation will be the same (0 and 1, respectively) for both groups. Herein lies the ingenuity. With this simple transformation, we now have a scale that we can set to have the same mean and standard deviation for any number of subgroups. Furthermore, it is a simple mathematical calculation to convert any *z* score to

have any other mean and standard deviation we choose. In the context of assessing intelligence, the obvious choice was the one that practitioners had become accustomed to. Hence, raw scores were transformed so that the mean for each age subsample of the standardization group was set to be 100, and the standard deviation was set to be around 15 (different test developers set slightly different standard deviations). This new IQ is referred to as the deviation IQ. It does not depend on a concept of mental age, and chronological age is used only for grouping.

An important caveat to this applies. The appropriateness of the linear transformation to *z* scores and the subsequent interpretations are premised on the assumption that the original raw scores fall along an equal interval scale. Equal interval scaling between scores is required if we are to make meaningful comparisons of differences. This has continued to be a major controversy in psychological assessment generally and one that is rarely questioned in clinical applications of IQ.

Contemporary Perspectives on IQ: A Cautionary Concluding Note

IQ is an aggregated score. To the extent that it is meaningful to aggregate scores across disparate tasks, such as Binet presented and as used in modern tests, such a single score is potentially appropriate. However, the current dominant theories suggest that intelligence is multifaceted and composed of distinct, though related, classes of ability. For instance, using sophisticated statistical techniques, McArdle, Ferrer-Caja, Hamagami, and Woodcock investigated the developmental trajectories of a range of different cognitive abilities. They interpreted significant deviations in the trajectories of these separate abilities from the *IQ-equivalent* trajectory to suggest that a description of the cognitive system with only a single factor is overly simplistic. Hence, although the notion of IQ has become deeply ingrained in modern language, great care is required with interpretation.

—Damian P. Birney and Steven E. Stemler

See also Intelligence Tests

Further Reading

- Binet, A., Simon, T., & Terman, L. M. (1980). *The development of intelligence in children* (1916 limited ed.). Nashville, TN: Williams Printing.
- Gould, S. J. (1996). *The mismeasure of man* (Revised and expanded ed.). New York: Penguin.
- Jensen, A. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger/Greenwood.
- McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, 38(1), 115–142.
- Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.
- Plucker, J. A. (Ed.). (2003). *Human intelligence: Historical influences, current controversies, teaching resources*. Retrieved from <http://www.indiana.edu/~intell>
- Stankov, L. (2000). The theory of fluid and crystallized intelligence: New findings and recent developments. *Learning and Individual Differences*, 12, 1–3.
- Sternberg, R. J. (1997). *Successful intelligence: How practical and creative intelligence determine success in life*. New York: Plume.
- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD: Williams & Wilkins.

INTELLIGENCE TESTS

An intelligence test is a structured situation designed to elicit information about the cognitive abilities of an individual. The test may be administered individually or in a group. Scores are usually reported on a scale in which 100 indicates average intelligence. Scores are scaled so that about the top 16% of the population will receive scores of 115 or above, the top 2.5% will receive scores of 130 or above, the bottom 16% will receive scores of 85 or below, and the bottom 2.5% will receive scores of 70 or below.

The typical intelligence test will have a variety of items designed to tap different aspects of the person's cognitive abilities. Some of the items may ask for specific pieces of information, such as how many years there are in a decade or how much change you would receive if you bought an article costing \$18.67 and

you gave the clerk a \$20 bill. Other questions might ask about objects missing or out of place in a picture; still others would be tests for memory, such as repeating a list of 5 digits that have been read, or tests of reasoning such as finding the right pattern piece to complete a design. In an individually administered test, the examiner asks each question, records the answer, and makes a judgment as to the answer's correctness or quality. Testing stops when the examinee has failed to answer a specified number of questions correctly. When the test is administered to a group, the questions are often in multiple-choice format, and responses are usually recorded by filling in bubbles on the answer sheet. Answers are compared to a key, so judgment as to correctness is avoided.

History of Intelligence Testing

Alfred Binet is generally given credit for creating the first modern intelligence test in 1905. In the 1908 version of his test, Binet introduced the idea of “mental level” as a way to express the cognitive ability of a child. The mental level of an item was the age at which the average child could solve that particular problem. An item that could be solved by the average child of age 7 or above, but not by a child of age 6, was given a mental level of 7 years. Items were grouped by mental level, and testing ended at the first level where a child could not answer any items correctly.

Henry Goddard popularized Binet's 1908 test in the United States. Several English-language versions of the Binet scale were quickly developed by Goddard and others. In 1916, Louis Terman published an American edition that came to be called the Stanford-Binet and soon replaced all competitors. This test popularized the term *intelligence quotient*, or IQ, because scores were expressed as the ratio of mental level or mental age, divided by actual or chronological age. A child who tested “at age” received an IQ of 1.00. This ratio came to be multiplied by 100 to remove the decimal point, resulting in a scale where the average IQ is 100, the reference point still in use.

During World War I, American psychologists under the leadership of Robert Yerkes produced two new group-administered tests for screening Army draftees:

a verbal form, Form Alpha, for those who could read English, and a nonverbal form, Form Beta, for the illiterate and those who did not speak English. Form Alpha had eight subtests and Form Beta had seven. Subtest scores were combined to produce a total IQ. After the war, the use of intelligence tests in schools, for college admissions, and in industry spread rapidly. Soon, several million tests, mostly of the group variety, were being administered each year.

In the 1930s, David Wechsler applied intelligence testing in the psychiatric clinic of Bellevue hospital. The Stanford-Binet had been developed for use with children and Wechsler needed a test for adults. He adapted for individual administration the tests from the Army testing program, grouped the tests into a verbal scale and a performance scale, and collected norms for adults. The resulting test battery was called the Wechsler-Bellevue Intelligence Scale. The verbal subtests were combined to produce a verbal IQ, or VIQ, and the nonverbal subtests yielded a performance IQ, or PIQ. The verbal and performance scores were combined to produce an overall description of cognitive ability called full-scale IQ, or FSIQ. At the same time, Wechsler rejected the IQ as the ratio of mental age to chronological age and introduced the scale that all intelligence tests use today, a scale with a mean of 100 and a standard deviation of 15.

Theories of Intelligence

Binet had viewed intelligence as a unitary concept that he equated with judgment. In 1904, Charles Spearman had suggested that all cognitive performances depended on a person's level of general cognitive ability, which he labeled *g*, and a specific ability required by that task and no other, called *s*. Spearman's "two-factor theory" (of *g* and *s*) agreed with Binet's ideas and his test, but it was rejected and harshly criticized by many American psychologists, particularly Edward L. Thorndike and his students.

A major factor in the debate was the nature of the battery of tests that each camp analyzed. Spearman usually had a small sample of subjects, each of whom had taken one test for each aspect of mental ability, much like those included in the Army testing

program. Thorndike usually had a large sample who had taken a much larger set of tests, several for each type of ability. When L. L. Thurstone developed the methods of factor analysis in the 1930s and analyzed a very large battery of tests, he found that there were about eight identifiable dimensions of cognitive ability that he called the Primary Mental Abilities. However, he also found that these primary mental abilities were all positively correlated. Spearman interpreted this finding that all mental abilities were positively correlated as vindication of his theory.

The debate over the structure of cognitive abilities continued through World War II. Thurstone's factor analytic results led to the development of batteries of tests to assess his ability factors as well as others. The most extreme of the multifactor theories was proposed by J. P. Guilford in the 1950s. Guilford's Structure of Intellect model eventually postulated 120 relatively independent abilities organized along the dimensions of Contents, Products, and Operations. Today, the main remaining parts of Guilford's theory are the concepts of divergent and convergent thinking. Divergent thinking is the ability to see nontraditional solutions to problems and is offered as a major component of creativity, whereas convergent thinking, the ability to extract a single correct solution to a problem, is seen as key to performance on most intelligence tests.

Development of CHC Theory and Related Tests

In the 1940s, Raymond B. Cattell identified two broad classes of intellectual functioning: the ability to solve new problems, which he labeled Fluid intelligence or *Gf*, and the fund of knowledge and information one had acquired from experience, which he called Crystallized intelligence or *Gc*. Starting in the 1960s, Cattell and his student, John Horn, expanded this theory to include eight or nine broad abilities. A massive factor reanalysis of more than 400 previous studies led John Carroll to offer a similar theory in 1993. Carroll's three-stratum theory postulated about 70 narrow, specific abilities that could be grouped into nine broad abilities, which in turn gave rise to a single general ability factor at the highest level. Cattell,

Carroll, and Horn agreed that they had arrived at essentially the same place, and their unified theory is now referred to as the Cattell-Horn-Carroll or CHC theory. CHC theory now forms the theoretical basis for almost all of the commercially available intelligence tests.

Reaction-Time Theory

Beginning about 1980, a new line of research into intelligence began in the work of Earl Hunt and Arthur Jensen. A century earlier, Sir Francis Galton had proposed that reaction time could be used to measure intelligence, but the primitive instruments available lacked sufficient accuracy to detect differences. With the advent of computers capable of accurately timing both stimulus presentation and response time, speed of neural processing became a potent area of research on intelligence. As research using these methods progressed, speed of information processing (called the chronometric approach to intelligence) and the extent and efficiency of working memory were found to correlate substantially with scores on traditional intelligence tests. The massive body of work by Jensen has been particularly influential. He and his colleagues have shown that the aspect of traditional intelligence tests that correlates most highly with processing speed/working memory is the single general ability factor most similar to Spearman's *g*.

One of the characteristics of processing speed that makes it attractive as a measure of intelligence is its relative freedom from cultural influences. One of the enduring criticisms of intelligence tests is that they are culturally biased, thereby resulting in lower scores for individuals who are not from the culture producing the test (generally upper middle-class Whites in the United States). Critics of intelligence testing claim that the tests measure exposure to the majority culture and therefore underpredict the success of members of minority groups. Various lines of evidence do not support these claims, but the claims continue nonetheless. Because information processing tasks can be designed to separate reaction time and movement speed from decision speed, relatively pure measures of an individual's ability to process information that are largely

independent of experience are possible. The fact that these measures correlate most highly with those traditional intelligence measures that are least influenced by culture, such as Raven's Progressive Matrices, suggests that information processing and fluid intelligence are similar.

Contemporary Intelligence Tests

A large number of intelligence tests have been developed over the years to measure various aspects of cognitive ability. Those that are offered for sale are reviewed in the *Mental Measurements Yearbooks*, published about every 2 years by the Buros Institute for Mental Measurements at the University of Nebraska. Some, such as the Stanford-Binet Fifth Edition and the Woodcock-Johnson Third Edition, are intended for use with the full range of ages from 2 years to more than 80, although the specific tests used at each age may differ. Others, such as the Wechsler Scales, have different tests for young children (the Wechsler Pre-School and Primary Scale of Intelligence), school-age children (the Wechsler Intelligence Scale for Children), and adults (the Wechsler Adult Intelligence Scale). Others, such as the Kaufman Ability Battery for Children, are only for a specific and limited age range, but almost all of them make an effort to measure some or all of the group factors in the CHC model. They all also offer a single index of general cognitive ability.

Intelligence or cognitive ability tests can be differentiated from achievement tests chiefly in the use to which the scores are put. Achievement tests measure the fund of knowledge one has acquired in a specific academic or occupational domain, such as language arts or mathematics. Achievement tests are postinstruction measures, and interpretation of the scores should normally be restricted to the domain and instructional experiences of interest. Identical items might appear on the quantitative or verbal ability subtests of an intelligence test battery, but in this context, the items are not linked to specific instructional objectives or curricula. Rather, they are seen as samples of larger domains of tasks, and the purpose of the testing is to assess potential for further learning. Thus, the

primary difference between achievement tests and some aspects of intelligence tests is the interpretation of the scores rather than their form or content.

—Robert M. Thorndike

See also Intelligence Quotient

Further Reading

- Flanagan, D. P., Genshaft, J. L., & Harrison, P. L. (1997). *Contemporary intellectual assessment: Theories, tests, and issues*. New York: Guilford.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Thorndike, R. M. (1990). *A century of ability testing*. Chicago: Riverside.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Buros Institute for Mental Measurements: <http://www.unl.edu/buros/>

Harcourt/PsychCorp: http://www.harcourt.com/bu_info/harcourt_assessment.html

Pearson Assessments: <http://www.pearsonassessments.com/index.htm>

Riverside Publishing: <http://www.riverpub.com/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Richardson, K. (2002). What IQ tests test. *Theory & Psychology*, 12(3), 283–314.

And isn't the title of this article the question that everyone asks? Kenneth Richardson contends that there is still little scientific agreement about how human intelligence should be described, whether IQ tests actually measure it, and if they don't, what they actually do measure. The controversies and debates that result are well known, and this paper brings together results and theory rarely considered in the IQ literature. It suggests that all of the population variance in IQ scores can be described in terms of a nexus of socio-cognitive-affective factors that differentially prepares individuals for the cognitive, affective, and performance demands of the test—in effect, that the test is a measure of social class background, and not one of the ability for complex cognition as such. The rest of the paper discusses

how such factors can explain the correlational evidence usually thought to validate IQ tests, including associations with educational attainments, occupational performance, and elementary cognitive tasks, as well as the intercorrelations among tests themselves.

INTERNAL EXTERNAL LOCUS OF CONTROL SCALE

The Internal External Locus of Control Scale (I-E Scale) is a researcher-constructed scale published by Julian B. Rotter in 1966. Social Learning Theory provides the theoretical background for which the scale provides a measure of generalized expectancy. The scale relates general expectancy as being either internally or externally controlled. An internally controlled person will perceive his or her destiny as controlled from within, whereas the externally controlled person perceives his or her destiny as controlled by outside forces such as chance, fate, or luck.

Rotter conceived the scale from earlier attempts by Phares and James to create a generalized expectancy scale in which they constructed their scales using the Likert format. Attempting to broaden the test, additional subscales were developed by Rotter for areas including achievement, affection, and general social and political attitudes. What resulted was a 100-item scale containing items comparing an internal belief to an external belief. In order to provide greater control for social desirability, the Likert format was discarded in favor of the forced-choice format. Using item analysis and factor analysis, this test was further reduced to 60 items by Liverant. Item analysis of the 60-item scale resulted in abandoning efforts to measure specific subareas. These items were then excluded. Additional items were also removed if they had a high correlation with the Marlowe-Crowne Social Desirability Scale, had an alternative endorsed more than 85% of the time, had no significant relationship with other items, or had little correlation with one of two criteria. The criteria used to provide

validity for the items involved a laboratory task and the coping behavior of tuberculosis patients. This reduced the scale to 23 items, to which six filler items were added for purposes of ambiguity, all in a forced-choice format. Two separate factor analyses determined that one factor accounted for most of the variance.

To score the test, a summation of external choices is made. High scores indicate externality and low scores indicate internality. The distribution of scores tends to be normal, and dividing groups about the median score is commonly used to indicate the split between externals and internals.

The I-E Scale can be used in a variety of situations. Since its conception, the I-E Scale has generated considerable interest and subsequent research. For example, externals have been associated with field dependence and internals with field independence. Interestingly, it has been posited that moderate internal scores should have some relationship to good adjustment. Accordingly, it has been reported that therapeutic outcomes are related to shifts in locus of control; that is, those who improve generally show a shift toward the internal end of the spectrum.

—John R. Reddon and Shane M. Whipple

See also Personality Tests

Further Reading

- Rotter, J. B. (1966). Generalized expectancies for the internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80, 1–28.
- Rotter, J. B. (1975). Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology*, 43, 56–67.
- Yalom, I. D. (1980). *Existential psychotherapy*. New York: Basic Books.

INTERNAL REVIEW BOARD

Internal Review Boards (IRBs), also known as Institutional Review Boards, are present at many

educational and health service institutions (e.g., universities and hospitals). The purpose of IRBs is to provide an objective and impartial review of research that is being conducted at the institution that involves human participants. The purpose of this review is to protect the welfare and well-being of those who are participating in the research.

History

The Nuremberg Code

The formal origin of human subject protections and the initiation of events leading to the formation of IRBs can be traced to the mid-20th century. During the Nuremberg Trials (1946–1949) that followed the end of World War II, it came to light that prisoners of war were required to take part as subjects in experimentation that sought to examine the effects of high altitude, drugs, experimental surgeries, poisons, and different types of explosive ordnance on human subjects. Many prisoners of war and civilians were required to take part in these activities against their will, and without any understanding of the procedures and their long-term effects. Furthermore, an examination of the research itself indicated that much of the experimentation was not scientifically justifiable. In 1947, concurrent with the Nuremberg Trials, 10 rules were set forth as essential principles in the conduct of research with human participants. These rules came to be known as the Nuremberg Code (1947/1996), and form the basis for the governance of human experimentation today. They represented the first codification of the principles of obtaining voluntary and informed consent from participants, the participants' right to withdraw from the research, the necessity of scientific justification and scientifically trained personnel in the conduct of the work, and the moral imperative to seek an experimental design that minimizes risk to participants.

Events Leading to the National Research Act of 1974

Shortly thereafter, the protection of human subjects became an important issue in U.S. research

communities. In the 1950s, the NIH established an internal ethics board for the review of intramural clinical research. In 1966, with concerns mounting from reviews of the ethics of clinical medicine, the U.S. Public Health Service (USPHS) issued a policy memo that mandated the formation of “review committees” and the requirement that research be reviewed from a human subjects protections perspective as a prerequisite for the receipt of federal funding for research. Clarifications and additions to this policy followed. In the early 1970s, however, public disclosure of clear ethical violations at the Willowbrook State School for the Retarded in New York and the USPHS Tuskegee studies resulted in the National Research Act of 1974, which established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The charge of the commission was to identify basic ethical principles to govern the conduct of biomedical and behavioral research involving human subjects, and the means by which these principles might be followed and enforced. The Commission produced The Belmont Report, which eventually evolved into the current federal code that governs the operation of IRBs in the United States.

Establishment of the Code of Federal Regulations

In the United States, the purpose, necessity, and operation of IRBs was established by law in 1980 by the passage of Title 45 Code of Federal Regulations Part 46, Protection of Human Subjects (45 CFR 46). These federal regulations grew out of the Belmont Report. With leadership from the Department of Health and Human Services, the Federal Policy for the Protection of Human Subjects or the “Common Rule” was published in 1991 and subsequently adopted by many federal departments and agencies.

Basis for IRB Review and Approval: The Risk/Benefit Ratio

IRBs seek to protect human subjects who are participating in institutional research primarily by

evaluating research protocols (summaries of the purposes and procedures of research) that are prepared by researchers prior to the initiation of the research itself. In this evaluation, the IRBs weigh the benefits of the research to society and to the individual participants against the risks of the research to the participants. This Risk/Benefit analysis initially derives from the Nuremberg Code and provides the fundamental framework for all IRB decisions. Legally, IRBs place human experimentation into three levels of risk. Risk may be defined in terms of the probability of either physical or psychological harm to the participants, and may take many forms. For example, subjects may be put at risk because of the side effects of a new experimental drug or medical procedure, or because of the inadvertent release of confidential personal information provided to researchers during the course of the study.

In the *no risk* category, procedures are judged to be completely innocuous and present no foreseeable risk. In the category of *minimal risk*, the procedures may present some level of risk to participants, but no more than that which is normally encountered in the participants’ everyday lives. If the IRB judges the research to be *greater than minimal risk*, then the IRB must closely consider the benefits of the research. If the risk to participants is judged to exceed the perceived benefit of the research, then the IRB should not provide institutional approval for the conduct of the research. If, however, benefit exceeds risk, then the study may be approved.

Operation of IRBs

The federal regulations governing IRB operations are quite specific in how the IRB is to be constituted and the criteria on which it must base its decisions. These federal regulations also specify the minimum standards for research oversight; local boards may impose more rigorous criteria, owing to standards or norms that may be prevalent in the surrounding community. Although IRBs are to meet a minimum of one time a year, most IRBs meet more often than that, usually monthly.

IRB Membership and Authority

The IRB must have at least five members. These members must be drawn from various backgrounds, and members must have sufficient experience and expertise to review research applications. IRBs may not consist entirely of men or entirely of women, and no IRB may consist entirely of members of one profession. The IRB must include at least one member whose primary concerns are in scientific areas, at least one member whose primary concerns are in nonscientific areas, and at least one member not affiliated with the institution, other than being a member of the IRB. The IRB generally has ultimate authority over the conduct of research with human participants at any institution. For example, although an institution may decide not to allow IRB-approved research to be carried out, the institution may not allow the conduct of research that has not been approved by the IRB.

Types of IRB Review

IRBs conduct two basic kinds of review prior to the initiation of research activity: *administrative* and *full board* review. After research has been approved, IRBs maintain oversight over research by conducting meaningful *continuing review*.

Administrative Review

Administrative review includes a determination of whether proposed research meets criteria for exemption or expedited review; in such cases, the review is not conducted by the full IRB, although the board is apprised of the outcomes of such administrative review, and any member can call for a project reviewed administratively to be brought before the full IRB at a convened meeting. Research activity may be deemed as *exempt* (i.e., not subject to the federal regulations, and thus not subject to IRB approval) if it falls into a set of predefined categories described in 45 CFR 46.101. However, individual researchers may not make judgments or determinations as to whether their

own research falls into this category: It remains with the IRB to verify that the exempt category is appropriate. Another type of administrative review that IRBs may use is *expedited review*; this is for certain kinds of minimal risk research that fall into one of eight categories described in 45 CFR 46.110. The Chair of the IRB may review such research or may designate one or more members of the IRB to provide the expedited review.

Full Board Review

Research that involves vulnerable populations, such as children in a noneducational setting, prisoners, pregnant women, or cognitively impaired individuals, or research that involves nonvulnerable populations in research of greater than minimal risk to participants, requires full board review. This means that the research application is at some point submitted to the entire board for its review and comment. Many IRBs use a primary reviewer system, which means that several (usually three) IRB members are asked to provide formal reviews of the project. The project is then put in a disposition category (e.g., approved, disapproved), and the full board is apprised of that disposition. Other board members may ask that a project be discussed at a convened meeting and may comment on applications in the primary reviewer system during those discussions.

Depending upon the IRB, the reviews that are reported to the research applicant may result in approval, a request for minor changes or clarification, or a call for more information or major changes in order to meet board approval. Applications that seem to pose an unacceptable risk to participants or that do not provide enough information to adequately review the application may require discussion by the convened IRB. Some IRBs may invite researchers to attend discussion of their projects by the IRB in order to provide additional insight to the research. Regardless of the type of review, IRBs are to notify researchers and the institution in writing of its decision to approve or disapprove the proposed

research activity, or of modifications required to secure IRB approval of the research activity. IRB approval of research activities is typically limited to a 1-year period.

Continuing Review

The fact that IRB approval is ephemeral allows for the conduct of meaningful *continuing review*, in which approvals may be reconsidered on an annual basis in light of changing local standards. In addition, IRBs monitor approved projects for the occurrence of adverse events within research studies, and they can suspend approval of a study if it is determined that the research activities may be causing harm to participants. Alternatively, if the evidence for the effectiveness of an experimental procedure is so robust that the project need not continue, such a decision would allow participants receiving placebo or control treatments to be given the actual treatment.

—David Hann and John Colombo

See also Ethical Issues in Testing; Ethical Principles in the Conduct of Research With Human Participants

Further Reading

- Beecher, H. (1966). Special article: Ethics and clinical research. *New England Journal of Medicine*, 274, 1354–1360.
- Belmont Report. (1979). *Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education, and Welfare.
- Citro, C. F., Ilgen, D. R., & Marrett, C. B. (Eds.). (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academies Press.
- Jones, J. H. (1993). *Bad blood: The Tuskegee syphilis experiment*. New York: Free Press. (Originally published in 1981)
- The Nuremberg Code. (1947/1996). *British Medical Journal*, 313, 1448.
- Rothman, D. J., & Rothman, S. M. (2005). *The Willowbrook wars: Bringing the mentally disabled into the community*. New Brunswick, NJ: Aldine Transaction.

INTERNATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The International Assessment of Educational Progress (IAEP) was created in 1988 for the purpose of international comparative studies. The IAEP collects and reports data on what students around the world can achieve academically.

The first IAEP was conducted to evaluate achievement in math and science of samples from students in five countries—Ireland, Korea, Spain, United Kingdom, and the United States—and four Canadian provinces. In 1991, the second assessment was conducted to assess math and science skills of 9- and 13-year-old students from as many as 20 countries.

The earlier version of the International Assessment of Educational Progress (IAEP-I) consisted of math and science questions derived from the National Assessment of Educational Progress (NAEP). The NAEP is an ongoing, congressionally mandated project designed to conduct national surveys of the educational attainments of students in the United States. Besides assessing math and science, the IAEP-I included questions about students' school experiences and attitudes. Measurement specialists from the United States and participating countries worked in translating and adapting the techniques used in the United States for the National Assessment of Educational Progress.

Like the previous version, the IAEP-II assessed math and science and included questions about students' backgrounds and home and school experiences. To assemble the tests, curriculum experts focused on common curriculum elements across countries to reach a consensus that produced the mathematics and science frameworks used in the development of the IAEP-II. However, some critics would argue that making comparisons of school quality across cultures is very difficult to interpret because each country has its own educational traditions and practices. For instance, the comparability of the student samples has been questioned because students in the United States remain in school for longer periods of time and are therefore part of the high

school test sample. In other countries, only the top students remain in school and are tested.

Despite the opinions of the critics, each of the countries that participated in these international comparative studies did so for its own reasons: to compare its results with those of other countries, to learn about the educational policies and practices of countries whose students seem to regularly achieve academic success, and to establish a baseline of data within its own country against which progress could be measured in the future.

—Romilia Domínguez de Ramírez

Further Reading

- Berliner, D. C. (1993, April). Mythology and the American system of education. *Phi Delta Kappan*, pp. 632–640.
- LaPointe, A. E., Mead, N. A., & Phillips, G. W. (1989). *A world of differences: An international assessment of mathematics and science*. New Jersey: Educational Testing Service.
- Medrich, E. A., & Griffith, J. E. (1992, January). *International mathematics and science assessment: What have we learned?* (National Center for Education Statistics, Research and Development Report). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement. Retrieved from <http://nces.ed.gov/pubs92/92011.pdf>
- International Comparative Studies in Education: Descriptions of selected large-scale assessments and case studies: http://books.nap.edu/html/icse/study_h.html
- Youth indicators 1993—School: Outcomes: <http://www.ed.gov/pubs/YouthIndicators/Outcomes.html>

INTERRATER RELIABILITY

Although there is no single definition of interrater reliability that is uniformly agreed upon in the statistical literature, there are generally two meanings associated with the term. The first meaning is represented by techniques that Stemler has classified as *consensus* estimates of interrater reliability. Consensus estimates of interrater reliability are used when the researcher is interested in examining the degree to which two or more independent raters can come to exact agreement about how to assign scores to observations

(or participants) based on a pre-established scoring protocol (or rubric).

The second meaning of the term is represented by techniques classified as *consistency* estimates of interrater reliability. Consistency estimates of interrater reliability are used when it is not critical for independent raters to agree exactly so long as the differences in how they apply the scoring rubric are systematic. For example, if one rater gave scores that were always two points lower than the second rater, the two raters would have low consensus estimates of interrater reliability, but high consistency estimates.

Consider the following data set in which two independent raters (Rater 1 and Rater 2) have assigned scores on a creativity scale (ranging from 1 to 5) to 15 students.

Student	Rater 1	Rater 2
1	1	2
2	1	1
3	1	2
4	1	2
5	2	3
6	2	3
7	2	3
8	2	4
9	3	4
10	3	3
11	3	4
12	3	5
13	4	5
14	4	4
15	4	5

Running the Crosstabs procedure in SPSS results in an output file with a table that provides a useful visual representation of the pattern of agreement between raters (see Figure 1).

Although a variety of statistical techniques may be used to compute consensus estimates of interrater reliability (e.g., Cohen's kappa, Jaccard's J), the most common consensus estimate of interrater reliability is the percent agreement statistic. Percent agreement is calculated by running a crosstab analysis and summing the values on the diagonals (3 in the example above) and dividing that value by the total number of

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Rater_1 * Rater_2	15	100.0%	0	.0%	15	100.0%

Rater_1 * Rater_2 Crosstabulation

		Rater_2					Total
		1.00	2.00	3.00	4.00	5.00	
Rater_1	1.00	1	3	0	0	0	4
	2.00	0	0	3	1	0	4
	3.00	0	0	1	2	1	4
	4.00	0	0	0	1	2	3
Total		1	3	4	4	3	15

Figure 1 Crosstabs Procedure Output

observations (15 in the example). This results in a percent agreement value of $\frac{3}{15} = 20\%$.

The most common consistency estimate of interrater reliability is the Pearson product-moment correlation coefficient; however, a variety of other consistency techniques also exist (e.g., Spearman's rho, Cronbach's alpha). The formula for computing the Pearson correlation is shown below:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

where

$\sum XY$ is the cross product of the raw scores;

$\sum X$ is the sum of all raw scores on variable X;

$\sum Y$ is the sum of all raw scores on variable Y;

$\sum X^2$ is the sum of all squared raw score values on variable X;

$\sum Y^2$ is the sum of all squared raw score values on variable Y;

$(\sum X)^2$ is the sum of all raw scores on variable X, squared;

$(\sum Y)^2$ is the sum of all raw scores on variable Y, squared;

n is the number of observations (participants) in the data set.

The SPSS output file above reveals that Raters 1 and 2 in the example have a low consensus estimate of interrater reliability (percent agreement = 20%) and a high consistency estimate ($r = .88$). Values greater than 70% or .70 are typically considered acceptable by the field.

If two raters have a high consensus estimate of interrater reliability, then the scores from the two raters may be treated as if they were equivalent and scores could be subsequently averaged,

or the scores of one rater randomly selected. By contrast, if two raters exhibit high consistency estimates of interrater reliability and low consensus estimates, the researcher must perform a correction to the data (e.g., add or subtract a constant) prior to summing or averaging the scores assigned by the two raters.

—Steven E. Stemler

See also Cohen's Kappa; Pearson Product-Moment Correlation Coefficient; Reliability Theory; Validity Theory

Further Reading

- Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, 16(3), 354–367.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved July 5, 2006, from <http://PAREonline.net/getvn.asp?v=9&n=4>
- Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1), 140–146.

Uebersax, J. (2002). *Statistical methods for rater agreement*. Retrieved August 9, 2002, from <http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Johnson, R. L., McDaniel, F., II, & Willeke, M. J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *American Journal of Evaluation*, 21(1), 65–80.

Portfolios and other open-ended assessments are increasingly incorporated into evaluations and testing programs. However, questions about the reliability of such assessments continue to be raised. After reviewing forces that may be leading to increased interest in and use of portfolio assessment, Robert Johnson and his colleagues investigated the **interrater reliability** of a portfolio assessment used in a small-scale program evaluation. Three types of portfolio scores were investigated—analytic, combined analytic (formed by summing across analytic scores), and holistic. The interrater reliability coefficient was highest for summed analytic scores ($r = .86$). Results indicate that at least three raters are required to obtain acceptable levels of reliability for holistic and individual analytic scores.

INTERVAL LEVEL OF MEASUREMENT

The interval level of measurement falls between the ordinal and ratio levels on the hierarchy of measurement and is considered the second “highest” or second most precise measurement scale. Interval scales are less exact than only ratio-level data. Interval data are more precise than nominal and ordinal data because the interval scale contains meaningful distances.

For instance, whereas in a nominal scale we can say only that certain traits or values belong to different mutually exclusive categories, and in an ordinal scale we can say only that a score or performance ranks above or below another one, with the interval scale, we can conclude *how much* higher one score is than another score. The unique characteristic of the scale that allows this determination of meaningful

distances (or the quantification of “how much”) is the equidistance of the intervals it contains.

To conceptualize the use of an interval scale, let’s explore an example that compares interval-level measurement to the two less precise measurement levels.

Five students took a 10-question true-or-false history quiz, and each question was worth one point. The students received the following scores:

Kevin—10 points

Jaime—6 points

Sara—5 points

Jen—3 points

Bill—0 points

In this example, the teacher could grade the students using a nominal scale and give each student a score of either “pass” or “fail” that corresponds to the student’s score. If the teacher decided to grade this way, Kevin and Jaime would receive the same grade, even though Kevin correctly answered four more questions than Jaime. Sara, Jen, and Bill would also earn equal grades of “fail” even though Sara got five more questions correct than did Bill and even though Sara’s score was only 1 point below Jaime’s.

The teacher could also decide to grade the quizzes using an ordinal scale. With this method, she would rank the students and give Kevin an “A” because he obtained the highest score, Jaime a “B,” Sara a “C,” Jen a “D,” and Bill an “F.”

However, the most precise way to grade the history quizzes would be to use an interval scale that corresponds directly to the number of questions the students answered correctly. Each interval along the scale would be worth the same amount: one point. Kevin would receive 10 points (100%), since he answered all questions correctly, Jaime would receive 6 points (60%), and so on. In this respect, we can say that Kevin’s score of 10 is 4 points higher than Jaime’s score of 6. An important point to remember, though, is that we are unable to say that Jaime has “twice” as much knowledge of history as Jen even though Jaime earned a 6 and Jen earned a 3. Because the interval scale has an arbitrary zero point (i.e., a score of 0 on the quiz does not indicate a complete

lack of knowledge about history), we cannot produce this ratio.

Using this interval scale for grading would highlight the most variability among the five students, telling us more about their knowledge of history than would nominal or ordinal grading methods. As Kevin would likely agree, the interval scale is probably the fairest grading method as well, because it clearly distinguishes the scores of 10 and 6 in a meaningful way. In the social sciences, especially because true ratio-level data are rare, interval-level data are desirable because of the amount of information they can provide.

—Kristin Rasmussen

See also Nominal Level of Measurement; Ordinal Level of Measurement; Ratio Level of Measurement

Further Reading

- Lane, D. (2003). *Levels of measurement*. Retrieved from <http://cnx.rice.edu/content/m10809/latest/>
- Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.
- Sirkin, R. M. (Ed.). (2005). *Statistics for the social sciences*. Thousand Oaks, CA: Sage.
- Trochim, W. M. K. (2002). *Levels of measurement*. Retrieved from <http://www.socialresearchmethods.net/kb/measlevl.htm>

IOWA TESTS OF BASIC SKILLS

The Iowa Tests of Basic Skills (ITBS), created at the University of Iowa, measures year-to-year growth of students' academic skills. Every year, thousands of students in states throughout the country take this battery of tests. The ITBS, formerly referred to as the Iowa Every-Pupil Tests, was first developed for high school students in 1931 by Hoover, Dunbar, and Frisbie. Shortly after, in 1935, this test was extended to the primary grades. The Iowa Every-Pupil Tests were renamed the Iowa Test of Basic Skills in 1955.

The ITBS was created as an achievement test to show the year-to-year progression of students. It also allows students, parents, and teachers a glimpse into an individual student's academic strengths and weaknesses, and it provides teachers with information

about the specific levels of their class compared to other tests in that district, state, and country to aid in teaching. Finally, the test is used to describe the student's developmental level, and the ITBS provides three types of scoring with each test: percentile ranks, grade equivalent, and standard scores.

The structure of the test is a basic multiple-choice test offering four possible answer options. The ITBS is taken in several subsections lasting approximately 30 minutes each, and the test authors suggest this 5.5-hour test be administered over 6 days. As with any test, revisions have been made over the years, with the 2001 version of the Iowa Test of Basic Skills being the most recent.

The ITBS contains questions in the areas of reading, language arts, mathematics, social studies, and science. The skills tested at the majority of levels include, but are not limited to, vocabulary, reading comprehension, language, an understanding of maps and graphs, categorizing, and math concepts and problems. Grades 3 through 8 also test the areas of history, economics, geography, and the life sciences. The scoring for the ITBS also provides subscores for each specific skill.

Below is an example of a potential math problem in the ITBS for Grade 5:

What is the value of m ?

$$6 + 2(m - 4) = 16$$

A. 8 B. 6 C. 2 D. 10

Currently, the ITBS test is published by Riverside Publishing.

—Sarah Peterson

See also Ability Tests; Achievement Tests

Further Reading

- Cummings, O. W. (1981). *Validation of a diagnostic interpretation technique for the Iowa Test of Basic Skills: Final report to the National Institute of Education*. Grant Wood Area Education Agency.
- Hieronymus, A. (1973). *Iowa Test of Basic Skills: Manual for administrators, supervisors, and counselors*. Chicago: Houghton Mifflin.

Hieronymus, A. (1986). *Iowa Test of Basic Skills: Forms G and H*. Chicago: Riverside.

Iowa Tests of Basic Skills information: <http://www.education.uiowa.edu/itp/itbs/index.htm>

IOWA TESTS OF EDUCATIONAL DEVELOPMENT

The most recent edition of the Iowa Tests of Educational Development (ITED) was published in 2001 (Form A) and 2003 (Form B) by Riverside Publishing Company. It is the 10th edition since the tests were first published in 1942. The tests provide objective information about high school (Grades 9–12) students' development in the skills that are the long-term goals of secondary education—skills such as

- Comprehending a wide variety of reading materials (both literary and informational)
- Solving quantitative problems
- Recognizing the essentials of correct and effective writing
- Critically analyzing discussions of social issues and reports on scientific matters
- Recognizing sound methods of scientific inquiry

The nine tests that comprise the ITED are Vocabulary, Reading Comprehension, Language: Revising Written Materials, Spelling, Mathematics: Concepts and Problem Solving, Computation, Analysis of Social Studies Materials, Analysis of Science Materials, and Sources of Information. About 4 hours of testing time are required to administer all nine tests. All questions are in the multiple-choice format and have four or five options each. Students mark their answer choices on a separate answer folder. The answer folder also contains space for students to respond to Interest Explorer (IE), an instrument that helps students identify career areas of interest to them. When students are exploring career options, the combined ITED/IE results can be a useful guidance tool.

The ITED were designed to fulfill three main educational purposes: (a) to obtain information that can be used to support instructional decisions (i.e., to

describe a student's developmental level within a test area and to identify a student's areas of relative strengths and weaknesses in the tested areas), (b) to examine the progress of grade groups as they pass through the school's curriculum from year to year, and (c) to provide information to students and their parents that helps them monitor a student's growth from grade to grade.

The ITED are standardized, norm-referenced achievement tests. The normative information was obtained in 2000 by administering the tests to thousands of students in Grades 9–12 across the nation. The normative data for the Iowa Test of Basic Skills (a battery of tests intended for students in Grades K–8) were gathered at the same time, and the two batteries are linked through their score scales. Thus, the two tests can be used to provide longitudinal achievement data for students in Grades K–12.

The ITED is developed at the University of Iowa. Over its long history, it has been known for its consistently rigorous measurement properties. A critical evaluation of the ITED can be found at the Buros Institute of Mental Measurements: Test Reviews Online (www.unl.edu/buros).

—Robert A. Forsyth

See also Ability Tests; Achievement Tests

Further Reading

Iowa Tests of Educational Development: Guide to research and development. (2003). Itasca, IL: Riverside.

Iowa Tests of Educational Development: Interpretive guide for teachers and counselors. (2001). Itasca, IL: Riverside.

IPSATIVE MEASURE

Data are ipsative if a given set of responses always sums to the same total. In practice, the term *ipsative* is used roughly as a synonym for “interdependent” and refers to some type of dependency among the variables measured on a survey, scale, test, or other measure. An example of data that are ipsative can be seen by asking respondents to choose between two items

that measure different psychological constructs. For example, a choice between “I am the life of the party” (extraversion) and “I am always prepared” (conscientiousness) will result in ipsative data because a choice of the extraversion item necessitates that the conscientiousness item is not chosen.

Many different properties of data collection and analysis exist that can create ipsative relationships between scale scores. For example, Likert scale data can be made to be ipsative by simply subtracting the grand mean of each individual’s scale scores (averaged across all scales) from each of his or her individual scale scores (i.e., ipsatized data). With this type of data, scores for each respondent will always sum to the same total across scales. However, there is no constraint on the variability of responses because respondents are free to choose any point on the scale without constraint. Data can also be ipsative via the properties of the item response format, such as rank-ordered scales, or through forced-choice responses from a set of items. With forced-choice and rank-order data, respondents are more constrained in their response options, and thus more interdependence exists in these types of ipsative data.

There are two primary types of interdependence that arise from ipsative data. The first type, covariance-level interdependence, relates to constraints that are placed on covariance matrices via the properties found in all types of ipsative data. Mathematically, (a) the sums of the columns, or rows, of an ipsative covariance matrix must equal zero; (b) the sums of the columns, or rows, of an ipsative intercorrelation matrix will equal zero if the ipsative variances are equal; (c) the average intercorrelations of ipsative variables have $-1/(m - 1)$ as a limiting value where m is the number of variables; (d) the sum of the covariances obtained between a criterion and a set of ipsative scores equals zero; and (e) the sum of ipsative validity coefficients will equal zero if the ipsative variances are equal. A second type of interdependence, item-level interdependence, occurs in rank-order and forced-choice scales because choosing any one item from a set is contingent upon the content of the other items.

Generally speaking, increasing the number of scales appearing on a survey will serve to lessen the

amount of covariance-level interdependence among constructs. Similarly, decreasing the percentage of measured scales that are used in analyses will also lessen the covariance-level interdependence. However, not using some scales in subsequent analyses will have no effect on the item-level interdependence of forced-choice and rank-order data. Although the issue remains controversial, these interdependencies can affect reliability estimates and factor analyses.

—Adam W. Meade

See also Measurement; Personality Tests

Further Reading

- Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouiser and spuriouiser: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153–162.
- Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, 77(4), 531–552.

ITEM AND TEST BIAS

Item and test bias have received much attention from the legal system, policymakers, test consumers, educational and psychological researchers, test developers, and the general public. This attention is well deserved because the essence of the issue is an ethical concern. *Bias* refers to differential validity across subgroups (e.g., males vs. females, minority vs. majority) and suggests that scores have different meanings for members of these subgroups. The Code of Professional Responsibilities in Educational Measurement states that test developers should make their products “as free as possible from bias due to characteristics irrelevant to the construct being measured, such as gender, ethnicity, race, socioeconomic status, disability, religion, age, or national origin” (Section 1.2a). However, there is no such thing as a “nonbiased test” or a test that is “fair” or “valid” for all subgroups under all conditions. This fact should not deter test developers from going to extensive

lengths to create instruments that are free of bias against intended subgroups.

To study whether performance may be influenced by factors specific to group membership (e.g., language, culture, gender), the psychometric properties of a test can be investigated for *invariance* (equality) across groups. The type of invariance investigation depends on the suspected nature of bias and can include a variety of methods to detect (a) differential item functioning (DIF), (b) factor structure invariance, and (c) differential prediction.

Item Bias and Differential Item Functioning

Although the terms *item bias* and *DIF* often are used interchangeably, DIF refers to differences in the statistical properties of an item between groups of examinees of equal ability. Two types of DIF can exist. *Uniform* DIF is a difference in item performance that is consistent across the ability distribution, whereas *nonuniform* DIF is a difference that is not consistent across the ability distribution (i.e., a group by ability-level interaction). Groups often are referred to as the *reference* (e.g., majority) and *focal* (e.g., minority or studied) groups. The concept of comparing groups of equal ability is a cardinal feature separating DIF from the traditional item bias detection methods. Traditional methods, because they do not control for ability differences, are affected by differences in the examinee group ability distributions. Overall ability differences may explain differential item performance, resulting in an item appearing to be, for example, more difficult when the examinees in the focal group are less able overall. *Impact* is a more appropriate term to refer to differences in item performance that can be explained by group ability differences. DIF detection methods “condition on” or control for ability, meaning that examinees are necessarily matched on ability; thus, only examinees of equal ability (e.g., total test score) in the reference and focal groups are compared.

Items that exhibit DIF threaten test score validity and may have serious consequences for groups as well as individuals, because correct responses are

determined by the trait claimed to be measured and factors specific to group membership. The most obvious consequence is the potential impact of DIF on the observed score distributions of specific groups. The less obvious consequence of DIF, yet critically important to the construct validity of a test, is its impact on the meaning and interpretation of test scores, even in the absence of mean score differences between groups. DIF items may cancel and result in similar score distributions across groups. However, when scores are composed of different items systematically scored as correct, it is invalid to infer that “equal” scores are comparable or have the same meaning. In fact, the Standards for Educational and Psychological Testing (Section 7.10) states that mean score differences are insufficient evidence of bias.

Many methods have been developed to detect DIF (e.g., SIBTEST, logistic regression, Mantel-Haenszel, item response theory methods). Generally, comparisons of methods reveal similar results given similar measurement models and testing conditions. Detailed descriptions of the various DIF methods can be found in Zumbo and Hubley’s 2003 review. However, the item response theory (IRT) likelihood ratio test is described to assist with understanding the general DIF detection process.

IRT DIF

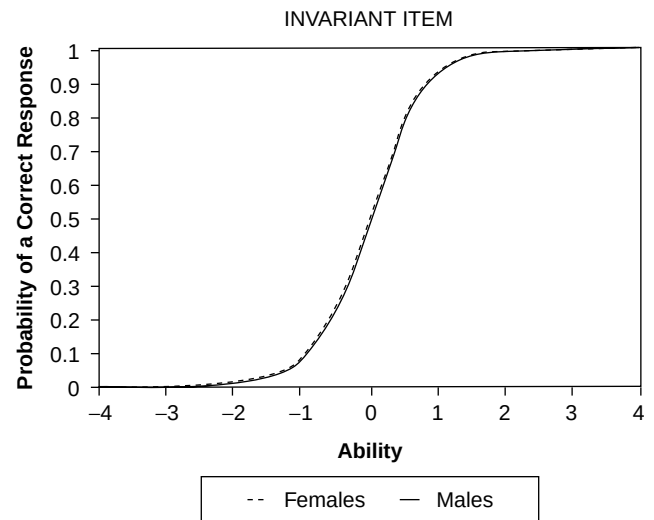
With IRT methods, the model conceptualizes ability as a latent trait compared to the typical observed test score. Formal tests of model-data fit provide greater confidence in the results, because a lack of model-data fit does not facilitate item parameter comparisons. Of particular interest to DIF are some of the IRT assumptions. First, IRT models assume invariance; that is, item parameters do not differ across groups unless DIF is present. In addition, IRT models generally assume unidimensionality (the test measures one dominant latent trait); however, the presence of DIF could signal multidimensionality in that construct-irrelevant factors related to group membership could influence item responses. Unfortunately, IRT methods require relatively large sample sizes (ranging from 100 to about 3,000), depending on the

model, than often are available for DIF studies with clinical tests.

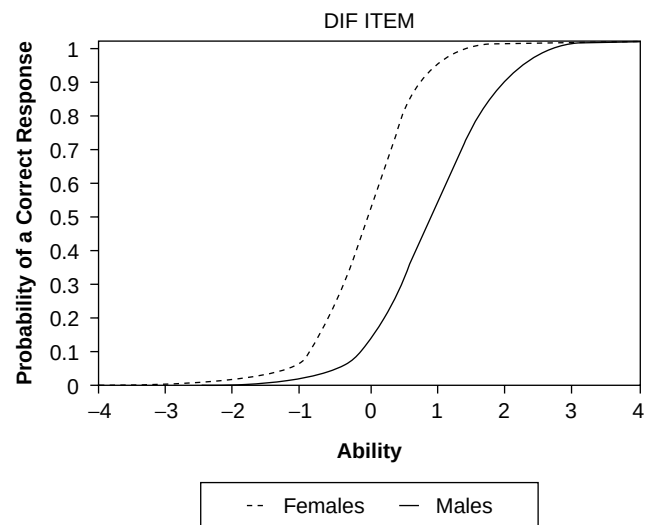
Figure 1 shows what are known as *trace lines*. The lines trace the probability (from zero to perfect probability) of a correct response to an item for persons from low to high ability. The first set of trace lines represents an invariant item (does not exhibit DIF), because the trace lines for males and females overlap. The second set of trace lines represents a DIF item. Notice that the lines are separated, and the probability of a correct response differs for males and females across all levels of ability. Following a multistep procedure, items found to be invariant across groups are used as a purified, or anchor, subtest to match examinees of equal ability, often based on an initial DIF screening with another DIF method (e.g., Mantel-Haenszel). *Purification* is recommended, for obvious theoretical reasons, to remove DIF items that might contaminate matching, although it is more labor intensive and thus more expensive. The items not included in the purified subtest are then individually tested for DIF using the following IRT likelihood ratio DIF detection method, similar to the method proposed by Thissen, Steinberg, and Wainer in 1993. Hypotheses are tested by comparing freely estimated difficulty or discrimination parameters to parameters constrained to be invariant across groups, using a likelihood ratio chi-square difference test. A nonsignificant test indicates no DIF. In addition to reporting effect sizes, exact probability values should be reported along with significance tests, given that numerous significance tests are reported. However, because power calculations are not yet available, it is recommended that a baseline estimate of the number of items that would exhibit DIF due to chance be obtained by dividing the sample in half at random and repeating all DIF analyses, and then calculating a *z* test of proportion differences.

Test Bias

Test bias refers to the differential validity of test scores. Investigations of test bias usually include studies of (a) unequal psychometric properties, including unequal factor structures; or (b) differential prediction of performance between groups.



(a) Trace Lines Displaying No DIF



(b) Trace Lines Displaying DIF

Figure 1 Item Trace Lines

Factor invariance suggests that test constructs are conceptualized and measured similarly across groups. Multisample confirmatory factor analysis can be used to test the invariance of factor structures, as described by Bollen in 1989, and allows for a chi-square test of model fit across groups. The general form of the theoretical model is tested for invariance to obtain model fit statistics across groups. If the general form does not fit across groups, test constructs are measured differently across the groups. If the model adequately

fits across groups, progressively more restrictive models are then tested for invariance. Three progressively more restricted models may be tested by adding one additional constrained matrix of (a) factor loadings, describing the relationships between the latent variable and an observed variable; (b) error variances; and (c) factor variances and covariances. Each element of a matrix can be individually tested for invariance to determine the sources of the lack of invariance.

Latent means structures (LMS) analysis is conducted when the measurement model is invariant. The group-measured variable mean is separated into two components: the intercept of the measured variable and the mean of the latent variable, where the intercept represents examinees' predicted subtest scores. Measured variable intercepts must be invariant to conclude that expected subtest score differences can be attributed to differences in the underlying ability and not potential systematic measurement bias, with one group systematically scoring higher or lower on a measured variable. If intercepts are invariant, group latent means are compared.

Multiple indicator, multiple cause modeling is a structural equation modeling procedure that can serve as an alternative to LMS to test for latent mean differences and potential measurement bias, and is especially useful for small samples. Specifically, groups are not divided, and thus, latent mean differences can be tested without the estimation of separate covariance matrices.

The model is composed of the following three components: (a) measurement model, (b) regression model, and (c) a "direct effects" path between the observed independent (group) and dependent variables (e.g., subscales). The measurement model represents the relationship between the observed and latent variables. The regression model consists of the effect of the dichotomous (0, 1) group indicator on the latent variables. A direct effect of the group variable to a subscale suggests that variance in the measure is explained by group membership in addition to the latent trait, thus indicating test-level bias. Such an effect is determined by a statistically significant improvement in model fit, based on the chi-square difference test.

Prediction bias, or the examination of differential predictive validity, is especially important when tests are used for placement and selection decisions. Prediction bias may occur when (a) validity coefficients significantly differ between groups, or (b) a test systematically under- or overestimates a criterion for a given group. That is, examinees from different subgroups, but with comparable predictor scores, obtain different scores on the criterion test. Differential prediction is investigated by comparing focal and reference groups' regression lines for criterion and predictor scores. A lack of differential prediction does not guarantee that a test will show item and test invariance. Moreover, criterion-related validity coefficients may be spuriously correlated because of systematic factors attributable to group membership. Thus, differential prediction should not be investigated without evidence of item and test invariance.

Current Issues and Future Directions

There has been much progress over the past two decades in the development and evaluation of bias detection methods. Several challenges, however, remain. First, much of the development to date has been on detection methods and less on the reasons for bias in applied settings. Second, uncertainty remains about how much bias is practically too much. That is, questions such as, What percentage or magnitude of item or test bias does it take to influence score-based decisions? and Does testing condition play a role—that is, paper-and-pencil scale, computer adaptive test, and test adaptations for special populations? deserve attention. Theoretical psychometric work has gone on to address this question, but more work needs to be done focusing on applied issues. Third, refinement of methods must continue. For instance, theoretically preferred methods (e.g., IRT methods) require rather large sample sizes. Continued evaluation and development of such preferred methods that allow use with sample sizes often seen with low-incidence populations (and other restrictive situations) are needed. Issues such as these, as well as others identified in the suggested readings, will keep both practitioners and

researchers quite busy over the next decade as the next generation of item and test bias methods are developed and implemented.

—Susan J. Maller, Brian F. French,
and Bruno D. Zumbo

See also Measurement

Further Reading

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response theory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from <http://educ.ubc.ca/faculty/zumbo/DIF/index.html>
- Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505–509). Thousand Oaks, CA: Sage.

ITEM RESPONSE THEORY

Item response theory (IRT) has many attractive features and advantages over classical test theory, which has contributed to its popularity in many measurement applications. Although IRT relies upon some strong assumptions, it is useful and practical in many situations, such as educational and psychological testing. IRT posits a probabilistic relationship between the response an examinee provides on a test item, or items, and some latent trait, such as ability or some personality trait. Although the topic of IRT is vast, this entry will attempt to discuss the assumptions of IRT, some of the more popular IRT models used for

dichotomously scored items, and some applications of the theory.

Popular IRT Models

Among the most popular IRT models used are those that are designed for applications where the test or test items measure a single underlying trait (unidimensional trait) by items that can be scored as 0 or 1 (dichotomous items). More complex models do exist for the cases where the items can be scored using multiple response categories (polytomous IRT models) and/or where the trait is multifaceted or multidimensional (multidimensional IRT models).

The most general of the models in this class is the three-parameter logistic model,

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability of a correct response to item I , given an ability level of θ . The item parameters are a_i , b_i , and c_i and refer to characteristics of the items themselves. The b parameter is often referred to as the item difficulty, and it is the point on the curve where the examinee has a probability of $(1 + c_i)/2$ of answering the item correctly. In the case where c_i is zero, that corresponds to the point where the examinee has a 50% chance of getting the item correct. The a parameter is commonly referred to as the item discrimination parameter, and it is the slope of the tangent line at the point on the θ scale equal to the b parameter. The c parameter is the pseudo-guessing parameter, or often, the guessing parameter, and is the height of the lower asymptote of the curve. This point provides the probability of a person of very low ability getting a correct response to the item. The curve generated by these item parameters is referred to as the item characteristic curve (ICC) or the item characteristic function (ICF). A graphical representation of an ICC with the corresponding parameters is presented in Figure 1.

Other popular IRT models are special cases of the more general three-parameter logistic model. The

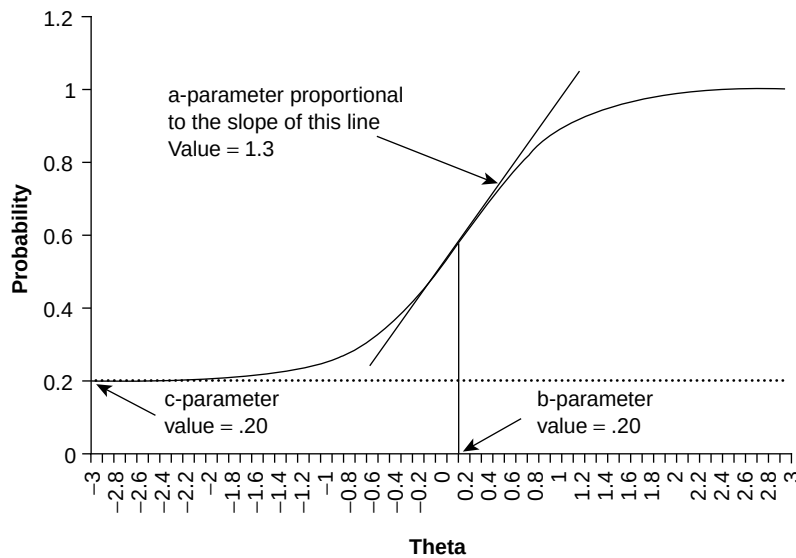


Figure 1 Representation of an ICC

two-parameter model is the case where the c parameter is set equal to zero. This model is used primarily in cases where guessing is not assumed to be a factor in the response to the item. The one-parameter model, or the Rasch model, is obtained when the c parameter is zero and the a parameter is set equal to 1 for all items. In this instance, items are assumed to be equally discriminating as well.

Assumptions and Features of IRT

IRT is often criticized for requiring strong assumptions that are difficult to attain in practice. Although it is true that IRT does rely upon strong assumptions, it has been applied successfully in many measurement applications. The assumptions are dependent on the type of IRT model chosen. In this case, we are referring to a particular class of IRT models: unidimensional models for dichotomous items.

In the case of unidimensional, dichotomous IRT models, which are the most commonly used models in measurement involving IRT, it is clear that there is an assumption that the underlying trait being measured is unidimensional. That is, there is one single trait that explains the behavior (responses) of the examinee to the test item(s). This assumption is testable, and techniques such as factor analysis, principal component analysis, and multidimensional scaling are often used

to test this assumption. Although most traits are typically not strictly unidimensional, in practice, this assumption is often relaxed to the case where the trait is “essentially” unidimensional, and this relaxation is usually not problematic in most applications.

The assumption of unidimensionality implies another assumption, namely, that of local item independence. Local item independence states that the response to one item is independent of the response to every other item, once the latent trait is taken into account. For example, in the case of a mathematics test, the responses to the items should depend only upon the mathematical

ability of the examinee, and no other factor. Once the mathematical ability of the examinee is taken into account, the responses to the items are independent of each other. This assumption likely would not hold in the case where the items required the examinee to read large amounts of text, because in that case, the reading ability of the examinee is likely to have an effect on the responses as well. In this instance, the items are not locally independent, and the trait being measured is not unidimensional because it consists of both math *and* reading ability.

The second major assumption of IRT models is that the model chosen adequately reflects the data, or that the model fits the data. There are several IRT models available for use, and the choice of the most appropriate model is critical for meaningful inferences to be drawn from the use of the model. Again, there are many techniques available for assessing the model-data fit, and the reader is referred to Hambleton, Swaminathan, and Rogers for more detail on this topic.

If the assumptions of IRT are met, however, there are many desirable measurement properties associated with the models. Among them is the invariance of the parameters. In IRT, the item parameters are inherent to the items themselves, and as such do not depend upon the sample of examinees responding to the items. Similarly, the level of the latent trait in the examinee is inherent to the examinee and is

not dependent upon the items that the examinee is administered. This feature of IRT allows for elegant methodologies for many measurement applications such as test construction, computer-based testing, equating, and identifying bias in items. Some of these applications will be discussed briefly below. Due to space limitations, the topics are not covered thoroughly, but rather are an attempt to illustrate how the properties of IRT can enhance certain assessment practices.

Some Applications of IRT

Test Construction

IRT can be a particularly valuable tool in constructing tests. The invariance of the item parameters allows parameters to be estimated using one sample, and then transferring the item characteristics to other, potentially unknown, samples of examinees. As such, once item characteristics are known, a test can be constructed with known difficulty, and other characteristics. Other test characteristics, such as the *test information*, can also be known a priori, which allows for the construction of a test with known precision, or reliability. To fully understand this topic, a discussion of test information is warranted.

Test information is determined by the items that comprise the test. Each item contains a certain amount of information about examinees along the ability scale. The item information for the three-parameter model is given by

$$I_i(\theta) = \frac{(1.7)^2 a_i^2 (1 - c_i)}{(c_i + e^{1.7a_i(\theta - b_i)})(1 + c_i + e^{-1.7a_i(\theta - b_i)})^2}.$$

As can be seen, the higher the a parameter, the more information that the item contains, and that the information is maximal for examinees with latent trait values near the value of the b parameter. Thus, information is not equal for all examinees, but rather different items provide differing amounts of information for each examinee; that is, information is a function of the ability of the examinee. As a result, Equation 2 is referred to as the item information function. The test

information function is formed by adding together the individual item information functions. Once the test information function is formed, the amount of information that exists at different points of ability can be assessed. The value of determining the amount of information is related to the precision of the test score; the more information that exists, the more precise the estimate of the latent trait. Because information is a function of θ , so too is the standard error. Therefore, a standard error can be determined for each level of the latent trait, rather than estimating one standard error that is common to every score. In fact, the standard error of the estimate of θ is given by the inverse of the information function; that is,

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

By understanding the item parameters and the item and test information functions, tests can be constructed by choosing items that best suit the purpose and features of the test, such as precision, and difficulty can be known *before* the test is administered. If it is important to have equal precision across the scale of the latent trait, then items can be selected so that the test information function is uniform. On the other hand, if the assessment is used for a pass/fail decision, it is necessary to minimize error at the cut score, and precision at either extreme is not necessary. In that case, items with maximum information at the cut score can be chosen, leading to a test with maximum information at the cut score, and hence maximum precision at the cut score. Therefore, depending on the purpose of the test, items can be selected in a purposeful manner to meet not only content requirements, but also statistically desirable characteristics.

Score Equating and Scale Linking

Because of the invariance of IRT item parameters, equating test scores is done fairly simply in the IRT context. Equating, or linking, is done when scores are to be compared on different versions of the same test. For example, in a test that is administered each year, the same test items may not be used every year, but it

is desirable to compare the scores on the two assessments. Therefore, the scores must be made comparable to each other before they can be compared. In the event that the tests are of identical difficulty, comparing scores can be done without any statistical adjustment. However, given the difficulty of achieving test forms that are identical in difficulty, it is unrealistic to assume that this case exists in practice. Therefore, the difficulties of the test forms must be made comparable first, through a statistical adjustment known as scale linking. Once the scales are made to be of the same difficulty, the scores can be adjusted in the process of score equating. Details of this process can be found in Kolen and Brennan.

The most difficult case of linking and equating occurs when the tests are different in difficulty and the examinee groups are different in ability. In this case, linking occurs by administering a common block of items to both groups of examinees. Because of the invariance property of the item parameters, the parameters should be the same in both groups of examinees. Granted, item parameter *estimates* cannot be assumed to be equivalent, but if the estimates are good, treating the estimates as the parameters is

usually adequate. Because of a scale indeterminacy issue in IRT, item parameters are invariant only up to a linear transformation, and the item parameters may not be identical if the two groups of examinees are different in ability. Thus, when differences are found, the linear transformation needed to put the parameters on the same scale can be found and the parameters can be adjusted to be equal. By applying this same transformation to the remaining item (and person) parameters, all parameters can be placed on the same scale, and scores can be equated. Without the invariance property of IRT, this process becomes much more complex.

—Lisa Keller

See also Measurement; Reliability Theory; Validity Theory

Further Reading

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices*. New York: Springer.

J

Nothing shocks me. I'm a scientist.

—Harrison Ford, as Indiana Jones

JACKSON, DOUGLAS N. (1929–2004)

Douglas N. Jackson made significant contributions to the field of psychometrics, particularly in the areas of construct validity, test construction, and response bias. He was also a prolific writer and test publisher with a commitment to free speech and academic independence.

He was born in Merrick, New York, and was dissuaded from pursuing an early interest in chemistry by career testing. The testing sparked a fascination with psychometrics and influenced the eventual decision to choose psychology as his profession. Graduating from Cornell University in 1951 with a bachelor of science degree in industrial relations and then from Purdue University in 1955 with a PhD in clinical psychology, he completed a postdoctoral internship at the Menninger Foundation. He was then at Pennsylvania State University and at Stanford University before taking a position at the University of Western Ontario, where he spent the rest of his career as senior professor of psychology.

A great deal of his work was focused on test construction and more specifically on method variance, which deals with the effects of such factors as format and wording on test results as opposed to the actual test-specific content that was to be measured. This work led to his concept that response styles influenced test outcomes and reflected traits of the respondent unrelated to those that were being measured. He was also a pioneer in the area of construct validity with an emphasis on theory, measurement model, and then data. He highlighted the importance of convergent and discriminant validity for the items of test scales. As a consequence, he advocated eliminating items that may be measuring more than one trait in an attempt to keep the scale as pure as possible. In addition to testing, he was well known for his work with person perception and determining how accurate our judgments of others are.

His interest in psychometrics did not end at the theoretical level. He constructed many tests in the field of personality and career planning, even founding his own publishing company (Sigma Assessment Systems, www.sigmaassessmentsystems.com) to produce and distribute his tests.

—John R. Reddon and Vincent R. Zalcik

See also Basic Personality Inventory; Jackson Personality Inventory–Revised; Jackson Vocational Interest Survey; Multidimensional Aptitude Battery

Further Reading

Goffin, R. D., & Helmes, E. (Eds.). (2000). *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy*. Boston: Kluwer.

Jackson Vocational Interest Survey online at <http://www.jvis.com/>

JACKSON PERSONALITY INVENTORY–REVISED

The Jackson Personality Inventory–Revised (JPI-R; publisher: Sigma Assessment Systems) was developed for assessing interpersonal, cognitive, and value domains of personality. Moreover, the JPI-R, which is used primarily with normal populations, also was designed to predict behaviors in a variety of contexts. The application of the JPI-R is most appropriate in career counseling and work settings that strive to improve person-environment fit; thus, the JPI-R frequently is used in schools, colleges, and career centers. In addition to applications in practice settings, the JPI-R is used in research settings to identify the relationship between certain personality constructs and behaviors.

The 300 items are presented in a true-false format that takes approximately 35–45 minutes to administer. The profile reports 15 content scales: complexity, breadth of interest, innovation, tolerance, empathy, anxiety, cooperativeness, sociability, social confidence, energy level, social astuteness, risk taking, organization, traditional values, and responsibility. A high score reflects the particular dimension the scale tries to measure; for instance, a high score on the Social Confidence scale indicates that an individual is confident in social situations. On the other hand, a low score on the Social Confidence scale suggests that a person will exhibit visible discomfort and embarrassment in most social situations. The 15 content scales were theoretically derived and constructed. They are classified into five clusters on the profile: Analytical, Emotional, Extroverted,

Opportunistic, and Dependable. The five clusters represent specific traits measured by the scales. For example, the Emotional cluster includes content scales such as empathy, anxiety, and cooperativeness. Factor analyses were applied to extract these five clusters.

The original JPI, published in 1976, was revised in 1994. The changes included renaming six scales, reorganizing profiles on the basis of newer research studies, removing the Infrequency scales, renorming the scales based on demographic samples that better represent the current population, adding and modifying current items, adding carbonless-form answer sheets, and adding new features to the manual.

The profile reports standardized scores that compare the test taker's score with a comparison norm group. The normative sample was collected from the general North American adult population. In addition, norm groups were collected for three main population groups: college students, blue-collar workers, and white-collar workers. Participants employed in a wide range of occupations were included in the blue-collar and white-collar samples. If an individual test taker is not a member of one of these three normative groups, the manual recommends that the individual's score be compared with the general population norm.

Evidence of reliability reported in the manual is limited to internal consistency coefficients, which ranged from $r = .78$ to $r = .93$. The manual also reported construct validity of the JPI-R. Studies that used multitrait-multimethod matrices showed evidence of convergence and discriminant validity at the factorial level. Overall, adequate evidence of validity and reliability has been collected for the JPI-R scales.

—W. Vanessa Lee and Jo-Ida C. Hansen

See also Personality Assessment Inventory; Personality Research Form; Personality Tests

Further Reading

Jackson, D. N. (1994). *Jackson Personality Inventory–Revised manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson Inventory information: <http://www.sigmaassessment.com/assessments/jpir.asp>

JACKSON VOCATIONAL INTEREST SURVEY

The Jackson Vocational Interest Survey (JVIS; publisher: Research Psychologists Press) was developed by Douglas N. Jackson to assist high school students, college students, and adults with career and educational planning. The survey is a self-report instrument that takes around 45 to 60 minutes to complete. The survey provides information on areas and patterns of vocational interests, together with indications of the similarity of a test taker's interests to those of individuals in various occupational and academic areas.

The JVIS profile includes 34 Basic Interests Scales. The definitions of these scales are based on a conceptualization of occupational preferences in terms of work roles and work styles. Work roles refer to relatively homogeneous samples of work activity, such as Teaching and Law. Work styles refer to preferences for working environments that require demonstration of behaviors such as Planfulness and Independence. The 34 Basic Interests Scales have been factor analyzed to identify 10 underlying dimensions. These dimensions, labeled as General Occupational Themes Scales on the JVIS profile, represent broad patterns of vocational interests, such as Expressive and Conventional.

In addition, an Academic Satisfaction Scale was empirically derived by comparing the average scores of high school and university students on the Basic Interests Scales. This scale may be useful in predicting degree of satisfaction in traditional academic settings. The profile also reports indices of the similarity between the test taker's Basic Interests Scales profile and the profiles of college students in 17 broad clusters of academic majors and employees in 32 job groups. The extended JVIS report contains further information related to career exploration.

Three useful administrative scales are the Unscorable Response, the Response Consistency Index, and the Infrequency Index. Internal consistency reliability of individual profiles can be determined on the basis of the Response Consistency Index. A combination of the Response Consistency Index and the

Infrequency Index provides evidence of validity for the profile scores.

The most recent normative sample was collected in 1999. The sample consisted of 1,750 males and 1,750 females from Canada and the United States. Among these 3,500 individuals, 2,380 were secondary students and 1,120 were university and college students and adults seeking career-interest assessment.

The internal consistency reliability (coefficients alpha) of the Basic Interests Scales based on the normative sample ranges from .54 to .88, with a median of .72. The internal consistency reliability for the 10 General Occupational Themes Scales ranges from .81 to .93, with a median of .88. The median test-retest reliability of the Basic Interests Scales and the General Occupational Themes Scales, based on a sample of university students, is .84 and .89, respectively.

One exemplary characteristic of the JVIS is the method used to construct the scales. Items on the JVIS were selected on the basis of a combination of construct and internal consistency criteria from a pool of more than 3,000 items. Another notable feature of the JVIS is the equal emphasis on the measurement of interests of women and men.

—Shuangmei (Christine) Zhou
and Jo-Ida C. Hansen

See also Armed Services Vocational Aptitude Battery; Career Assessment Inventory; Career Development Inventory; Career Maturity Inventory

Further Reading

Juni, S., & Koenig, E. J. (1982). Contingency validity as a requirement in forced-choice item construction: A critique of the Jackson Vocational Interest Survey. *Measurement and Evaluation in Guidance*, 14(4), 202–207.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

The *Journal of the American Statistical Association* (JASA) is the first serial publication of the American

Statistical Association (ASA), founded in 1839, and it remains the flagship publication of the association.

JASA began as *Publications of the American Statistical Association*, a series of tracts on statistical topics. The first appeared in March 1888 and contained only one 44-page article, "Statistics of Water Power Employed in Manufacturing in the United States," by George F. Swain. The next issue (vol. 1, nos. 2–3, June and September 1888) became more journal-like in character: It contained two articles by different authors, the second one a history of the U.S. Census, including an extensive bibliography and a key to U.S. Census publications. Subsequent issues grew rapidly in terms of the numbers of articles (most considerably shorter) and the scope of topics considered. The series continued through 16 volumes comprising 128 issues from 1888 to 1919. At that point, the series was renamed *Quarterly Publications of the American Statistical Association*, but only one volume (vol. 17, 1920–1921) was published under that title.

Volume 18 (1922–1923, issues 137–144) is the first that bears the title *Journal of the American Statistical Association*. That volume contained 70 full-length articles, 30 short notes, 63 book reviews, and 28 reviews of reports. Beginning in 1924, various supplements began to appear sporadically: proceedings of meetings and ASA handbooks that contained the constitution and bylaws, membership directory, and an index of past issues of *JASA*.

Francis A. Walker, the ASA president in the late 1880s, is credited with providing the impetus to publish a journal. The early volumes were edited by ASA secretary Davis R. Dewey, who served until 1908. The next ASA secretary was Carroll W. Doten, who apparently also edited the journal. Subsequent editors were William F. Ogburn (1920–1925), Frank A. Ross (1926–1934), Frederick W. Stephan (1935–1940), again Ross (1941–1945), and William G. Cochran (1945–1950). Starting in 1968, separate editors were appointed for applications and methodology papers (in addition to the review editor), and beginning in 1970 (vol. 65), the research articles were divided into two sections, titled "Applications" (later "Applications and Case Studies") and "Theory and Methods." The same

basic organization remains to the present day: three editors and separate boards of associate editors.

Today, *JASA* is universally considered to be one of the top nonspecialized journals in statistics. In terms of total citations in the science literature through 2003, the ISI Web of Knowledge ranks *JASA* at the top of all journals in the mathematical sciences. When journals in behavioral sciences, psychology, and education are included, *JASA*'s total citations rank behind those of *Psychological Bulletin* and *Animal Behavior* and about equal with *Psychological Review*'s.

—Russell V. Lenth

See also American Statistical Association; *Journal of Modern Applied Statistical Methods*; *Journal of Statistics Education*

Further Reading

ISI Web of Knowledge: [http://www.isiwebofknowledge.com/Journal of the American Statistical Association](http://www.isiwebofknowledge.com/Journal%20of%20the%20American%20Statistical%20Association) and its predecessors are archived on the JSTOR Web site: www.jstor.org *Journal of the American Statistical Association* and other ASA publications information: www.amstat.org

JOURNAL OF MODERN APPLIED STATISTICAL METHODS

The *Journal of Modern Applied Statistical Methods* (*JMASM*) was incorporated in 2000 and published its first issue in 2002. The founding editor is Shlomo S. Sawilowsky of the College of Education at Wayne State University, Detroit, Michigan.

The appearance of *JMASM* continues a long tradition of journal activity in Sawilowsky's academic genealogy. Excluding copious ad hoc editorial reviews, some examples include Gottfried Wilhelm Leibniz (1646–1716), cofounder of *Acta Eruditorum* in 1682 and founding editor of *Monatliche Auszug* in 1700; Joseph Liouville (1809–1882), founding editor of *Journal de Mathématiques Pures et Appliquées* in 1836; Eugène Catalan (1814–1894), founding editor of *Nouvelle Correspondance Mathématique* in 1874; Jules Tannery (1848–1910), coeditor of

Bulletin des Sciences Mathématiques beginning in 1876; and David Lee Hanson, associate editor of *Annals of Mathematical Statistics/Annals of Statistics* beginning in 1967 and the *Annals of Probability* in 1973.

JMASM is an independent journal. It is not supported by any professional or scholarly society or organization. It is a peer-reviewed print and “open access” electronic journal (<http://tbf.coe.wayne.edu/jmasm>). Open access means the journal is based on a funding model that does not charge readers or institutions for electronic access, and the electronic version is released simultaneously with the print version. It is published twice a year (May and November). Each issue contains about 295 pages. There have been more than 5,750 downloads for each of the first nine issues of the journal.

More than 250 universities and colleges worldwide receive the print journal, and more than 350 list the electronic version in their library. *JMASM* is a core journal in the Current Index to Statistics and is also indexed in Elsevier Bibliographic Database, EMBASE, Compendex, Geobase, Scopus, and ScienceDirect.

The inaugural editorial team of *JMASM* includes associate editors Bruno D. Zumbo of the University of British Columbia, Canada, and Harvey J. Keselman of the University of Manitoba, Canada, and assistant editors Alan Klockars of the University of Washington, Todd C. Headrick of the University of Southern Illinois–Carbondale, and Vance W. Berger of the National Cancer Institute, Bethesda, Maryland. The international editorial board represents about 50 universities and institutions, with about 85% of the members from the United States or Canada and the remaining from Greece, India, Israel, Italy, Jordan, Malaysia, Saudi Arabia, and the United Kingdom.

The primary purposes of *JMASM* are to promote the following:

1. The development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods
2. The development or study of nonparametric, robust, permutation, exact, and approximate randomization methods

3. Applications of computer programming related to statistical algorithms, pseudorandom number generators, simulation techniques, and self-contained executable code (e.g., Fortran) to carry out new or interesting statistical methods
4. Applied problems in statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation in all disciplines of systematic inquiry

Articles from renowned scholars appear in the Invited section. To date, these authors include Ralph D’Agostino, Sr.; James Algina; Peter Bentler; R. Clifford Blair; Robert Boik; Walt Brainerd; William J. Conover; C. Mitchell Dayton; Philip I. Good; Gregory R. Hancock; Harvey J. Keselman; Thomas R. Knapp; George Marsaglia; S. James Press; Pranab K. Sen; Ronald C. Serlin; Juliet P. Shaffer; Judith M. Tanur; Neil H. Timm; and Rand R. Wilcox.

The main section contains about 16 regular articles. Following are Brief Reports, Statistical Software and Applications Review, and Statistical Algorithms and Code, with about three articles in each. An innovative and important section, Early Scholars, provides journal space for doctoral students and is peer reviewed by the same.

There is an occasional section, compiled by the editor, called Statistical Pronouncements, which attempts to capture the philosophy, insight, wit, humor, and even the inevitable faux pas of mathematicians and statisticians. Other occasional sections include invited debates and interviews with mathematicians and statisticians. The concluding section contains both paid and pro bono advertisements from vendors of statistical software and products and many pro bono publico announcements pertaining to statistics and research journals and to professional and academic societies.

The mailing address is *JMASM*, PO Box 48023, Oak Park, MI 48237. Address e-mail not related to submissions, including requests for advertising requirements, to jmasm@edstat.coe.wayne.edu.

—Shlomo S. Sawilowsky

See also Journal of the American Statistical Association

Further Reading

Sawilowsky, S. S. (2004). A conversation with R. Clifford Blair on the occasion of his retirement. *Journal of Modern Applied Statistical Methods*, 3(2), 518–566.

JOURNAL OF STATISTICS EDUCATION

The *Journal of Statistics Education (JSE)* is a rigorously refereed electronic journal on teaching statistics. Its goal is to provide useful and interesting information, ideas, materials, software, and data sets to teachers of statistics at all levels. The intended audience includes both members of university statistics departments and faculty who teach statistics in other disciplines, such as mathematics, psychology, or sociology. The journal also publishes articles of interest to teachers of statistics at the primary or secondary level or in the workplace. It is available free of charge to any interested reader.

JSE has published articles on diverse topics, including assessment of students' understanding and attitudes, teaching inference and experimental design, distance learning, curriculum guidelines, cooperative learning, and using activities and projects. Regular features of *JSE* are "Data Sets and Stories" and "Teaching Bits: A Resource for Teachers of Statistics." Articles in the "Data Sets and Stories" department describe interesting data sets and their pedagogical features. The data sets themselves may be downloaded for analysis or use with students. "Teaching Bits" summarizes articles from newspapers, magazines, and other publications that can be used as examples in the statistics classroom.

The idea of establishing a journal on statistics education arose during a strategic planning effort in the Department of Statistics at North Carolina State University (NCSU) in 1991. We felt that a lack of publication outlets for work on teaching statistics made it difficult for statistics educators to exchange ideas and also discouraged statisticians from investing time in the scholarship of teaching. We were intrigued

by the possibility of publishing innovative content—interactive teaching materials, dynamic graphics, video, or sound—by using an electronic medium. At a planning workshop with 22 participants held at NCSU in 1992, the decision was made to establish a refereed electronic journal on teaching statistics.

The table of contents of the inaugural issue of *JSE* was sent by e-mail to 403 subscribers in July 1993. The first issue contained plain-text articles that were accessible by e-mail or gopher. The technological capabilities of the journal increased rapidly; the second issue contained the first graphic, the third issue was distributed via the World Wide Web, and the fourth issue included a brief animation. *JSE* was supported initially by the Department of Statistics at NCSU and a grant from the Fund for the Improvement of Postsecondary Education, U.S. Department of Education, and later by a grant from the Exxon Education Foundation. In 1999, *JSE* became an official journal of the American Statistical Association—the association's first electronic journal and its first journal devoted to statistics education. The journal now resides on the Web site of the American Statistical Association and is published in HTML format.

The first three editors of *JSE* were E. Jacquelin Dietz (1993–2000), Thomas Short (2001–2003), and W. Robert Stephenson (2004–2006). An international editorial board sets policy and handles the review process. Articles that pass an initial screening by the editor are sent to a member of the editorial board, who in turn sends the article to two additional reviewers. Refereeing is double blind. Refereeing criteria include quality of exposition, originality or novelty, and usefulness to teachers of statistics.

—E. Jacquelin Dietz

See also American Statistical Association; *Journal of the American Statistical Association*; *Journal of Modern Applied Statistical Methods*

Further Reading

Journal of Statistics Education home page: <http://www.amstat.org/publications/jse/>

K

Everywhere is walking distance if you have the time.

—Steven Wright

k-MEANS CLUSTER ANALYSIS

Many different cluster analysis methods are available to the researcher for classifying observations and variables. Most of these methods fall into one of two categories: hierarchical and nonhierarchical. The nonhierarchical methods are occasionally referred to as *iterative partitioning* methods. Hierarchical methods usually begin the clustering process by forming a matrix of similarities between entries. Clusters are formed by putting together those entries that are the most similar. This method is an agglomerative process that yields the results in the form of hierarchical trees or denograms. The nonhierarchical methods of iterative partitioning usually begin with an arbitrary classification, and through an iterative revision process, they attempt to find a classification that minimizes the within-cluster variation or equivalently maximizes the between-cluster variation.

Each of the two categories of method has its appropriate domain of applications and corresponding advantages and disadvantages. The nonhierarchical method appears to be especially appropriate in the reduction of large databases, in the analyses of group similarities, in nonlinear predictions, and in estimation

of multivariate distributions for high-dimensional metric data often found in the social sciences.

The purpose of this summary is to describe the algorithm for one particular method of nonhierarchical clustering. The method is known as *k-means*. Computer programs such as SPSS and SAS are set up to perform this type of clustering. The algorithm presented here is the one developed by James B. MacQueen. Also, we will discuss proposed enhancements to the algorithm that would provide the user of this method better information about the cluster solution.

The *k-means* algorithm was originally developed as a method of computing an optimal partitioning in the sense of within-class variance of an n -dimensional population on the basis of a sample. The *k-means* procedure is faster than most nonhierarchical methods, but it does not, in general, converge to a global optimum partition. There are, however, special cases for which convergence to an optimal partition can be achieved. There does not seem to be any feasible, general method that will *always* yield an optimum cluster solution except in one dimension. The *k-means* procedure appears to give clusters that are reasonably efficient in a within-class variance sense. The efficiency is dependent to some extent on the researchers' intuition

in picking the “correct” number of clusters (k) corroborated by mathematical analysis and practical computational experience.

Features of the Clustering Algorithm

To sort m data units into k clusters, the following steps are used:

1. The first k data units are taken as the first k clusters with one point in each cluster.
2. The following (or remaining) $m - k$ data units are assigned (one after another) to one of the k clusters on the basis of the shortest distance between the data unit and the centroid (mean) of the cluster. The distance can be computed using a number of formulas. The most popular is the Euclidean distance squared:

$$D = \sum_{i=1}^n (X_{ijk} - \bar{X}_{ik})^2.$$

This formula is the squared Euclidean distance between the centroid of cluster k and the j th data unit in that cluster.

After each assignment, the centroid is recomputed (updated) for that cluster, gaining a member. In situations where there are an extremely large number of cases, this step may be omitted in order to speed up the computations.

3. After the assignment of each data unit to one of the k clusters has been completed, the cluster centroids are taken as fixed seed points, and each data unit is again assigned to one of several clusters according to the closest distance to a centroid. After all data units have been reassigned, the centroid for each cluster is recomputed and stored for possible usage in future reclassifications.

Step 3 can be performed a specified number of times to help refine the classification of data units to clusters, or until there is no further reduction in the within-class variance. As with the updating of the cluster centroids, the reclassification step can be omitted in order to obtain a quicker solution.

A Needed Feature

In order to make the k -means clustering routine more useful than just a simple data reduction method, an additional feature is suggested here to help the researcher find “better” clusters.

A Randomization Test for Clusters

A feature needed in clustering programs is the capability for approximate (nonparametric) permutation or randomization tests. (NCSS 2004 [Number Cruncher Statistical Systems] can do this.) The values of any selected variables or any subgroup of variables can be put in random order to form a new sample in which the relationship between the selected variables of groups is completely randomized. For example, in two dimensions, the original sample $(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$ becomes $(X_{i_1}, Y_{j_1}), (X_{i_2}, Y_{j_2}) \dots (X_{i_n}, Y_{j_n})$, where $(i_1, i_2, \dots, i_n)$ and $(j_1, j_2, \dots, j_n)$ are independent random permutations of the integers 1, 2, 3, $\dots$, n . For each such randomization requested, the computer program is applied to the transformed data set as in the first instance.

A number of applications of this feature are possible. For example, to search for a better clustering, the whole data set is put in a random order. Because the within-cluster variance is related to the starting order, it may be possible to obtain a clustering with lower variance by such a search. Experience shows, roughly, that a 5% to 15% decrease in the variance may be expected from the worst to the best of a half-dozen such randomized starting positions.

Another use is to employ clustering with the randomization to perform an approximate nonparametric permutation test for association among all the variables. The idea is that the within-cluster variance of the clusters tends to be lower when the variables are related, and this circumstance is true even if the relationship is highly nonlinear. Under the hypothesis of independence, the original data set is considered a randomization. Hence, in 19 randomizations, for example, if the original data have a lower within-cluster variance than the within-cluster variance from the randomizations, rejection of the hypothesis of independence at the 5% level is valid. This decision

would be true for any measure of association. The test that can be applied to blocks of variables equally well, also provides significance tests for a sort of nonlinear method of canonical correlation. For example, all of the health variables could be randomized against all of the socioeconomic variables for a global nonlinear test of their association. Whenever complex statistical operations are performed, and whenever there is practically no hope of obtaining analytical distributions of the test statistics under a reasonable null hypothesis, the distribution-free character and ease of interpretation of these approximate randomization tests are extremely valuable.

Using the Computer

The data in the entry's tables were taken from the NEA Higher Education Advocate. The data consist of the average state faculty salaries from public 2-year and 4-year institutions from 2002–2003 to 2003–2004. Included are data on whether the state has

collective bargaining, and salary trends. The higher the trend number, the greater the ratio of faculty salaries to a four-person household income. Those with collective bargaining received a score of “1,” those without received a “2.”

The analysis called for three clusters. From the results given using SPSS, it appears that there are five states in which institutions of higher learning are distinct from the other 44 states. These have collective bargaining, higher trends, and higher salaries. Among the 44 states, 29 of them are in the lower salary classification, usually with no collective bargaining and with a lower level trend. There are 15 in the middle classification. These have a higher salary than the other 29, as well as collective bargaining. These 15 states, however, have a higher trend value for 2-year institutions than for 4-year institutions.

—Howard B. Lee

See also Distance; Metric Multidimensional Scaling

Table 1 State Faculty Salary Average and Grade for Public 2- and 4-Year Institutions, 2002–03 to 2003–04

State	Two-Year	Trend2	Four-Year	Trend4	Barg
AK	62,220	2	55,865	0	1
AL	43,829	3	56,122	3	2
AR	37,873	2	50,313	3	2
AZ	58,799	4	68,410	4	1
CA	70,305	4	78,168	4	1
CO	42,137	0	61,461	0	1
CT	59,729	2	73,402	1	1
DE	53,773	1	76,762	2	1
FL	47,306	3	63,391	4	1
GA	43,609	1	61,773	2	2
HI	52,506	1	65,832	1	1
IA	42,663	1	69,378	4	1
ID	41,988	2	51,125	1	2
IL	56,984	4	63,188	1	1
IN	41,821	1	63,277	2	2
KS	43,163	1	60,840	2	1
KY	44,428	2	58,220	3	1
LA	41,049	3	56,255	3	2
MA	55,574	2	67,076	1	1
MD	55,357	0	67,780	0	1
ME	44,745	3	56,215	2	1
MI	65,895	4	69,351	3	1
MN	54,285	0	67,915	0	1

(Continued)

MO	47,010	1	56,335	0	1
MS	42,595	4	52,275	4	2
MT	37,419	2	53,141	3	1
NC	37,906	0	62,267	4	2
ND	37,282	1	50,308	1	2
NE	40,775	0	61,893	1	1
NH	41,906	0	66,716	0	1
NJ	62,795	3	77,462	2	1
NM	41,224	2	57,978	4	1
NV	51,508	2	70,304	2	1
NY	59,421	4	67,474	3	1
OH	50,642	2	66,655	3	1
OK	40,689	2	52,798	2	2
OR	51,719	3	56,900	1	1
PA	54,213	4	69,441	3	1
RI	52,688	3	68,317	3	1
SC	40,498	1	58,918	3	2
SD	38,981	1	50,859	0	1
TN	41,841	2	58,493	4	2
TX	46,190	2	61,194	2	2
UT	41,746	0	57,385	0	2
VA	45,912	0	66,161	1	2
WA	48,153	2	63,240	1	1
WI	61,199	4	65,470	2	1
WV	40,497	4	51,934	4	2
WY	44,273	2	61,721	2	2

(Continued)

(Continued)

Output from SPSS**Quick Cluster****Initial Cluster Centers**

	Cluster		
	1	2	3
twoyear	62220.00	70305.00	37282.00
trend2	2.00	4.00	1.00
fouryear	55865.00	78168.00	50308.00
trend4	.00	4.00	1.00
barg	1.00	1.00	2.00

Iteration History^a

Iteration	Change in Cluster Centers		
	1	2	3
1	12102.413	8413.128	9282.466
2	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 23723.218.

Cluster Membership

Case Number	State	Cluster	Distance
1	AK	1	12102.413
2	AL	3	2718.484
3	AR	3	8984.476
4	AZ	1	4716.247
5	CA	2	8413.128
6	CO	3	3228.451
7	CT	2	3212.826
8	DE	2	8896.817
9	FL	3	7317.664
10	GA	3	3842.429
11	HI	1	2523.389
12	IA	3	11158.810
13	ID	3	7108.768
14	IL	1	3104.953
15	IN	3	5052.980
16	KS	3	2809.855
17	KY	3	2312.345
18	LA	3	2246.956
19	MA	1	1585.596
20	MD	1	2214.985
21	ME	3	3314.222
22	MI	2	6615.874
23	MN	1	2436.690

(Continued)

24	MO	3	5249.309
25	MS	3	5976.871
26	MT	3	6927.012
27	NC	3	5830.755
28	ND	3	9282.466
29	NE	3	3898.183
30	NH	3	8485.971
31	NJ	2	2450.891
32	NM	3	927.333
33	NV	1	5876.247
34	NY	1	4788.818
35	OH	1	4503.335
36	OK	3	5618.767
37	OR	1	9296.145
38	PA	1	3933.102
39	RI	1	3585.893
40	SC	3	1756.891
41	SD	3	8012.276
42	TN	3	378.491
43	TX	3	5036.842
44	UT	3	924.737
45	VA	3	8790.402
46	WA	1	7256.323
47	WI	1	6182.320
48	WV	3	6503.291
49	WY	3	4101.558

Final Cluster Centers

	Cluster		
	1	2	3
twoyear	55017.87	62499.40	42115.69
trend2	2.47	2.80	1.59
fouryear	65591.13	75029.00	58232.62
trend4	1.67	2.40	2.21
barg	1.00	1.00	1.59

Distances between Final Cluster Centers

Cluster	1	2	3
1		12043.532	14853.076
2	12043.532		26412.384
3	14853.076	26412.384	

Number of Cases in each Cluster

Cluster	1	15.000
	2	5.000
	3	29.000
Valid		49.000
Missing		.000

Further Reading

- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley and Los Angeles: University of California Press.
- Silver, N. C., & Hittner, J. B. (1998). *A guidebook of statistical software for the social and behavioral sciences*. Needham Heights, MA: Allyn & Bacon.

Clustering algorithm tutorial: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

KAUFMAN ASSESSMENT BATTERY FOR CHILDREN

The Kaufman Assessment Battery for Children, Second Edition (KABC-II), published by American Guidance Service, is designed to measure cognitive processing and aptitude in children ages 3 through 18. It can be administered nonverbally to children who have hearing impairments, limited English proficiency, or language disorders, and it can be adapted for use with children from diverse cultural backgrounds. The instrument can be used in combination with other assessment tools to identify children with cognitive and academic deficits such as mental retardation or learning disabilities; children with cognitive processing difficulties; and children who have cognitive giftedness. Interpretation of KABC-II scores can also help identify a child's cognitive strengths and weaknesses.

The KABC-II may be administered in 30–90 minutes, depending upon the age and ability of the child. Altogether, the instrument contains 18 subtests. For each age group, 5–10 of the subtests are identified as “core” subtests and are recommended to obtain an optimal measure of cognitive functioning across five main areas (or scales). The Sequential Processing Scale contains short-term memory tasks; for example, children listen to a series of simple words and recite them in order. The Simultaneous Processing Scale focuses on children's ability to process visual stimuli. Tasks include using a two-dimensional diagram to construct a tower of blocks. The Learning Scale

measures long-term memory by presenting novel learning tasks and measuring the child's ability to remember and apply new knowledge. The Planning Scale includes fluid reasoning tasks; for example, children are required to solve problems logically. Finally, the Knowledge Scale measures vocabulary and other verbal skills by having the child complete tasks such as naming pictures. The results from these scales also produce a global scale, which represents a measure of the child's overall cognitive functioning.

This instrument was updated in 2004 and includes significant changes from the first edition. The age range has been expanded to include children up to age 18, and stimulus items have been revised to engage very young children as well as to challenge adolescents. The first edition of the KABC was firmly grounded in Luria's simultaneous-sequential cognitive processing model, which separates intelligence into two parts: (a) processing that leads to finding interrelationships and (b) processing that allows for arranging information logically. The KABC-II preserves Luria's model, but also introduces a “dual theoretical model” by incorporating the Cattell-Horn-Carroll (CHC) theory of cognitive processing. The CHC theory also separates intelligence into two parts: (a) fluid intelligence, which incorporates both pieces of Luria's model, and (b) crystallized intelligence, which reflects acquired knowledge. This change has resulted in several new subtests, as well as the addition of the Learning, Planning, and Knowledge Scales.

—Carrie R. Ball

See also Stanford-Binet Intelligence Test

Further Reading

- Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of K-ABC-II assessment*. Circle Pines, MN: American Guidance Service.
- Samuda, R. J. (1998). *Advances in cross-cultural assessment*. Thousand Oaks, CA: Sage.
- Alan S. Kaufman and Nadeen L. Kaufman (test developers) biographical information: <http://www.mhhe.com/mayfield/pub/psychtesting/profiles/karfmann.htm>
- American Guidance Service: <http://www.agsnet.com>

KENDALL RANK CORRELATION

The Kendall rank correlation coefficient evaluates the degree of similarity between two sets of ranks given to the same set of objects. This coefficient depends upon the number of *inversions* of pairs of objects that would be needed to transform one rank order into the other. In order to do so, each rank order is represented by the set of all pairs of objects (e.g., $[a,b]$ and $[b,a]$ are the two pairs representing the objects a and b), and a value of 1 or 0 is assigned to this pair when its order corresponds or does not correspond to the way these two objects were ordered. This coding schema provides a set of binary values that is then used to compute a Pearson correlation coefficient.

Notations and Definition

Let S be a set of N objects,

$$S = \{a, b, \dots, x, y\}. \quad (1)$$

When the order of the elements of the set is taken into account, we obtain an ordered set that can also be represented by the rank order given to the objects of the set. For example, with the following set of $N = 4$ objects,

$$S = \{a, b, c, d\}, \quad (2)$$

the ordered set $\mathcal{O}_1 = [a, c, b, d]$ gives the ranks $\mathcal{R}_1 = [1, 3, 2, 4]$. An ordered set on N objects can be decomposed into $\frac{1}{2}N(N-1)$ ordered pairs. For example, $\mathcal{O}_1$ is composed of the following six ordered pairs:

$$\mathcal{P}_1 = \{[a, c], [a, b], [a, d], [c, b], [c, d], [b, d]\}. \quad (3)$$

In order to compare two ordered sets (on the same set of objects), the approach of Kendall is to count the number of different pairs between two ordered sets. This number gives a distance between these sets called the *symmetric difference distance* (the symmetric difference is a set operation that associates to two sets the set of elements that belongs to only one set).

The symmetric difference distance between two sets of ordered pairs $\mathcal{P}_1$ and $\mathcal{P}_2$ is denoted $d_\Delta(\mathcal{P}_1, \mathcal{P}_2)$.

The Kendall coefficient of correlation is obtained by normalizing the symmetric difference such that it will take values between -1 and $+1$, with -1 corresponding to the largest possible distance (obtained when one order is the exact reverse of the other order) and $+1$ corresponding to the smallest possible distance (equal to 0, obtained when both orders are identical). Taking into account that the maximum number of pairs that can differ between two sets with $\frac{1}{2}N(N-1)$ elements is equal to $N(N-1)$, this gives the following formula for the Kendall rank correlation coefficient:

$$\begin{aligned} \tau &= \frac{\frac{1}{2}N(N-1) - d_\Delta(\mathcal{P}_1, \mathcal{P}_2)}{\frac{1}{2}N(N-1)} \\ &= 1 - \frac{2 \times [d_\Delta(\mathcal{P}_1, \mathcal{P}_2)]}{N(N-1)}. \end{aligned} \quad (4)$$

How should the Kendall coefficient be interpreted? Because τ is based upon counting the number of different pairs between two ordered sets, its interpretation can be framed in a probabilistic context. Specifically, for a pair of objects taken at random, τ can be interpreted as the difference between the probability of these objects being in the same order [denoted $P(\text{same})$] and the probability of these objects being in a different order [denoted $P(\text{different})$]. Formally, we have

$$\tau = P(\text{same}) - P(\text{different}). \quad (5)$$

An Example

Suppose that two experts order four wines called $\{a, b, c, d\}$. The first expert gives the following order: $\mathcal{O}_1 = [a, c, b, d]$, which corresponds to the following ranks: $\mathcal{R}_1 = [1, 3, 2, 4]$; and the second expert orders the wines as $\mathcal{O}_2 = [a, c, d, b]$, which corresponds to the following ranks: $\mathcal{R}_2 = [1, 4, 2, 3]$. The order given by the first expert is composed of the following six ordered pairs:

$$\mathcal{P}_1 = \{[a, c], [a, b], [a, d], [c, b], [c, d], [b, d]\}. \quad (6)$$

The order given by the second expert is composed of the following six ordered pairs:

$$\mathcal{P}_2 = \{[a, c], [a, b], [a, d], [c, b], [c, d], [d, b]\}. \quad (7)$$

The set of pairs that are in only one set of ordered pairs is

$$\{[b, d], [d, b]\}, \quad (8)$$

which gives a value of $d_\Delta(\mathcal{P}_1, \mathcal{P}_2) = 2$. With this value of the symmetric difference distance, we compute the value of the Kendall rank correlation coefficient between the order given by these two experts:

$$\begin{aligned} \tau &= 1 - \frac{2 \times [d_\Delta(\mathcal{P}_1, \mathcal{P}_2)]}{N(N-1)} \\ &= 1 - \frac{2 \times 2}{12} = 1 - \frac{1}{3} \approx .67. \end{aligned} \quad (9)$$

This large value of τ indicates that the two experts strongly agree on their evaluation of the wines (in fact, they agree about everything but *one* pair). The obvious question now is to assess if such a large value could have been obtained by chance or can be considered evidence for a real agreement between the experts. This question is addressed in the next section.

Significance Test

The Kendall correlation coefficient depends only on the order of the pairs, and it can always be computed

assuming that one of the rank orders serves as a reference point (e.g., with $N = 4$ elements, we assume arbitrarily that the first order is equal to 1234). Therefore, with two rank orders provided on N objects, there are $N!$ different possible outcomes (each corresponding to a given possible order) to consider for computing the sampling distribution of τ . As an illustration, Table 1 shows all the $N! = 4 \times 3 \times 2 = 24$ possible rank orders for a set of $N = 4$ objects along with its value of τ with the “canonical order” (i.e., 1234). From this table, we can compute the probability p associated with each possible value of τ . For example, we find that the p value associated with a one-tailed test for a value of $\tau = \frac{2}{3}$ is equal to

$$\begin{aligned} p &= P\left(\tau \geq \frac{2}{3}\right) = \frac{\text{Number of } \tau \geq \frac{2}{3}}{\text{Total number of } \tau} \\ &= \frac{4}{24} = .17. \end{aligned} \quad (10)$$

We can also find from Table 1 that in order to reject the null hypothesis at the alpha level $\alpha = .05$, we need to have (with a one-tailed test) perfect agreement or perfect disagreement (here, $p = \frac{1}{24} = .0417$).

For our example of the wine experts, we found that τ was equal to .67; this value is smaller than the critical value of +1 (as given by Table 2), and therefore, we cannot reject the null hypothesis, and we cannot conclude that the experts displayed a significant agreement in their ordering of the wines.

The computation of the sampling distribution is always theoretically possible because it is finite. But

Table 1 The Set of All Possible Rank Orders for $N = 4$, Along With Their Correlation With the “Canonical” Order 1234

Rank Orders																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3	3
3	4	2	4	2	3	3	4	1	4	1	3	2	4	1	4	1	2	2	3	1	3	1	2	2
4	3	4	2	3	2	4	3	4	1	3	1	4	2	4	1	2	1	3	2	3	1	2	1	1
τ	1	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$-\frac{1}{3}$	$\frac{1}{3}$	0	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{2}{3}$	-1

Table 2 Critical Values for τ for $\alpha = .05$ and $\alpha = .01$

	<i>N = Size of Data Set</i>						
	4	5	6	7	8	9	10
$\alpha = .05$	1	.8000	.7333	.6190	.5714	.5000	.4667
$\alpha = .01$	–	1	.8667	.8095	.7143	.6667	.6000

this requires computing $N!$ coefficients of correlation, and therefore, it becomes practically impossible to implement these computations for even moderately large values of N . This problem, however, is not as drastic as it seems because the sampling distribution of τ converges toward a normal distribution (the convergence is satisfactory for values of N larger than 10), with a mean of 0 and a variance equal to

$$\sigma_{\tau}^2 = \frac{2(2N + 5)}{9N(N - 1)}. \quad (11)$$

Therefore, for N larger than 10, a null hypothesis test can be performed by transforming τ into a Z value as

$$Z_{\tau} = \frac{\tau}{\sigma_{\tau}} = \frac{\tau}{\sqrt{\frac{2(2N + 5)}{9N(N - 1)}}}. \quad (12)$$

This Z value is normally distributed with a mean of 0 and a standard deviation of 1.

For example, suppose that we have two experts rank ordering two sets of 11 wines. The first expert gives the following rank order:

$$\mathcal{R}_1 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11],$$

and the second expert gives the following rank order:

$$\mathcal{R}_2 = [1, 3, 4, 5, 7, 8, 2, 9, 10, 6, 11].$$

With these orders, we find a value of $\tau = .6727$. When it is transformed into a Z value, we obtain

$$Z_{\tau} = \frac{.6727}{\sqrt{\frac{2(2 \times 10 + 5)}{9 \times 11 \times 10}}} = \frac{.6727}{\sqrt{\frac{54}{990}}} \approx 2.88. \quad (13)$$

This value of $Z = 2.88$ is large enough to reject the null hypothesis at the $\alpha = .01$ level, and therefore, we can conclude that the experts are showing a significant agreement between their evaluations of the set of wines.

Kendall and Pearson Coefficients of Correlation

The Kendall coefficient of correlation can also be interpreted as a standard coefficient of correlation computed between two sets of $N(N - 1)$ binary values, where each set represents all the possible pairs obtained from N objects; it assigns a value of 1 when a pair is present in the order and 0 if not.

Extensions of the Kendall Coefficient

Because the Kendall rank order coefficient relies on a set distance, it can be generalized easily to other combinatoric structures such as weak orders, partial orders, or partitions. In all cases, the idea is similar: first compute the symmetric difference distances between the two sets of pairs representing the binary relation, and then normalize this distance so that it will take values between -1 and $+1$.

—Hervé Abdi

See also Distance; z Scores

Further Reading

- Degenne, A. (1972). *Techniques ordinales en analyse des données*. Paris: Hachette.
- Hays, W. L. (1973). *Statistics*. New York: Holt, Rinehart and Winston.
- Kendall, M. G. (1955). *Rank correlation methods*. New York: Hafner.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

KINETIC FAMILY DRAWING TEST

The Kinetic Family Drawing (KFD) is a projective test used to understand and assess the perspectives of children and adolescents on their families.

Interpretation of the KFD is based on the “projective hypothesis,” which is an assumption that an individual, when drawing a picture on a blank page, will project his or her thoughts, concerns, conflicts, needs, motivations, and frustrations into the picture. According to one of the originators of the KFD, Robert C. Burns, the KFD allows us to see the self as it is reflected and expressed in the family; it enables the young person to depict the family as a functioning, active unit and allows us to see the child’s impressions of these dynamic interactions among family members. Administration of the test is simple, requiring the individual to “draw a picture of your whole family, including yourself, doing something” on a sheet of plain white paper.

According to Burns, scoring of the picture uses four major categories: Actions; Distances, Barriers, and Positions; Physical Characteristics of the Figures; and Styles. Actions refers to the content or theme of the drawing and it may symbolize cooperation, communication, masochism, narcissism, nurturance, sadism, or tension. Physical Characteristics represents formal aspects of the drawings (e.g., inclusion/exclusion and size of essential body parts, the relative sizes of the figures, and facial expressions). Distances, Barriers, and Positions refers to the barriers between figures, the direction faced by each figure, and the distances between figures. Styles refers to the organization of the figures on the page and includes descriptors such as Compartmentalization (intentional separation of family figures using lines); Edging (placement of all figures on the perimeter of the paper); Encapsulation (encapsulating one or more figures by lines or objects); and Underlining individual figures (lines immediately below a standing individual or individuals). The various categories are scored and then interpreted. For example, Burns suggested that tension and instability in the family may be demonstrated by compartmentalization, subgroups as opposed to a united family, or barriers between people. Feelings of isolation or rejection may be evidenced by no face on the self, a distorted self-figure, or orientation of one or both parents away from the self.

A review of the reliability and validity studies of the KFD is complicated by the fact that there have

been several scoring systems proposed for the KFD, sometimes with additional scoring variables. Thus, the KFD remains a clinical instrument with inadequate norms and questionable validity. Numerous authors have criticized the somewhat arbitrary interpretations of drawing variables based upon factors that have not been validated empirically. Therefore, the clinical skill of the interpreter is very important in the interpretation of the drawings, possibly accounting for the disparate results found by researchers who are not first and foremost clinicians. Most researchers have treated the scoring variables independently of each other, finding nonsignificant or contradictory results, whereas a qualitative, integrative, holistic scoring system has shown more promise. Clinicians tend to use the KFD as an icebreaker with clients and as a starting point for conversations about family. It is added to a range of clinical techniques and psychometric tests to provide more information about the individual’s perspectives on the self in the family.

Robert C. Burns has written most prolifically about the KFD, but a large number of books are available about the KFD by other authors, applying the test to different populations of children. The KFD has no copyright and is in the public domain.

—Fran Vertue

Further Reading

- Burns, R. C. (1982). *Self-growth in families: Kinetic Family Drawings (KFD) research and application*. New York: Bruner/Mazel.
- Handler, L., & Habenicht, D. (1994). The Kinetic Family Drawing technique: A review of the literature. *Journal of Personality Assessment*, 62, 440–464.
- Knoff, H. M., & Prout, H. T. (1985). *Kinetic drawing system for family and school: A handbook*. Los Angeles: Western Psychological Services.
- Niolon, R. (2003). *Notes on projective drawings*. Retrieved on August 20, 2005, from http://www.psychpage.com/projective/proj_draw_notes.html
- Tharinger, D. J., & Stark, K. (1990). A qualitative versus quantitative approach to evaluating the Draw-A-Person and Kinetic Family Drawing: A study of mood- and anxiety-disordered children. *Psychological Assessment*, 2, 365–375.
- Wegmann, P., & Lusebrink, V. B. (2000). Kinetic Family Drawing scoring method for cross-cultural studies. *The Arts in Psychotherapy*, 27, 179–190.

KINGSTON STANDARDIZED COGNITIVE ASSESSMENT

The Kingston Standardized Cognitive Assessment–Revised (KSCA-R) is a Canadian-developed dementia screen. This measure was designed to fill an important gap between rating scales that are too brief, nonspecific, or narrowly focused, such as the Mini Mental State Examination (MMSE), and more comprehensive but lengthy and expensive neuropsychological assessments. The KSCA was designed to be administered by clinicians who do not necessarily have specialty training in psychometrics or test interpretation.

Like the original KSCA, the Revised KSCA, published in 2003, generates a total score and three sub-total scores: Memory, Language, and Visual-Motor. The revised version takes approximately 20 to 30 minutes to administer. The most notable content change in the revised version is the replacement of the four-word memory task with a list-learning task in the Memory subscale. In addition, the new Assessment Form has been improved to facilitate scoring and interpretation. A separate manual provides detailed administration and scoring procedures as well as the new normative data. Furthermore, an interpretive section has been added to the end of each subtest section in the manual.

The KSCA-R is normed on a community living outpatient sample that has met the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*, and National Institute of Neurological and Communication Disorders and Stroke criteria for probable Alzheimer's disease and a sample of normal, healthy, community dwelling elderly. The inclusion of the new word list memory task has dramatically enhanced the KSCA-R's ability to detect early-stage dementia, long before the MMSE can do so.

The KSCA-R has been compared to traditional neuropsychological measures [i.e., the Wechsler Adult Intelligence Scale-III (WAIS-III) and the Wechsler Memory Scale-III]. As predicted, strong positive correlations were obtained (e.g., KSCA-R Total Score $\times$ WAIS-III Verbal IQ = 0.85; KSCA-R Memory Subtotal $\times$ WAIS-III Immediate Memory Index = 0.76).

The KSCA-R has a new set of “score analysis pages” that uses a combination of flow charts and tables to guide the user through the process of converting the scores to percentiles and descriptive ranges. These pages also help the user make clinical decisions.

The KSCA-R is made available by the authors free of charge and may be obtained by contacting them at hopkinsr@post.queensu.ca. A training video has also been developed.

—Lindy Kilik, Robert Hopkins,
and Duncan Day

Further Reading

- Hopkins, R., Kilik, L., Day, D., Rows, C., & Hamilton, P. (2004). The Revised Kingston Standardized Cognitive Assessment. *International Journal of Geriatric Psychiatry*, 19, 320–326.
- Hopkins, R., Kilik, L., Day, D., Rows, C., & Hamilton, P. (2005). The Brief Kingston Standardized Cognitive Assessment–Revised. *International Journal of Geriatric Psychiatry*, 20, 227–231.

KOLMOGOROV-SMIRNOV TEST FOR ONE SAMPLE

Generally, Kolmogorov-Smirnov tests are aimed at testing the hypothesis that two or more distributions are identical. The one-sample version tests the hypothesis that observations were sampled from a specified distribution. For example, one could test the hypothesis that observations arise from a normal distribution having mean 3 and standard deviation 6. Or one could test the hypothesis that sampling is from a chi-squared distribution with 6 degrees of freedom. So the one-sample version does not test the hypothesis that observations follow some normal distribution having some unknown mean and variance; rather, it can be used to test the hypothesis that observations follow a precisely specified distribution. However, a simple extension of the method can be used to test the hypothesis that observations follow a normal distribution with unknown mean and variance. The two-sample version tests the hypothesis that two unknown

distributions are identical. Certain advances make it a potentially useful method for getting a detailed description of how distributions differ that goes beyond any technique based on a single measure of location.

For the one-sample version considered here, let $F_0(x) = P(X \leq x)$ be the known (specified) distribution, and let $X_1, \dots, X_n$ be a random sample of size n from the unknown distribution $F_1(x)$. Letting $I_{X_i \leq x} = 1$ if $X_i \leq x$, otherwise $I_{X_i \leq x} = 0$, F_1 is estimated with

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x},$$

the proportion of observations less than or equal to x .

The two-sided version is designed to test

$$H_0 : F_1(x) = F_0(x), \text{ all } x$$

versus

$$H_0 : F_1(x) \neq F_0(x), \text{ for at least one } x.$$

The test statistic is based on what is sometimes called the Kolmogorov distance, which is just the maximum absolute difference between the two distributions under consideration. More formally, the test statistic is

$$T = \max |F_1(X_i) - F_0(X_i)|,$$

the maximum being taken over all $i = 1, \dots, n$.

There are one-sided versions of the test as well. The first tests

$$H_0 : F_1(x) \geq F_0(x), \text{ all } x$$

versus

$$H_1 : F_1(x) < F_0(x), \text{ for at least one } x.$$

The test statistic is

$$T^+ = \max |F_0(X_i) - F_1(X_i)|.$$

The other one-sided version tests

$$H_0 : F_1(x) \leq F_0(x), \text{ all } x$$

versus

$$H_1 : F_1(x) > F_0(x), \text{ for at least one } x.$$

The test statistic is

$$T^- = \max |F_1(X_i) - F_0(X_i)|.$$

For all three versions, the null hypothesis is rejected if the test statistic is sufficiently large.

Using a recursive algorithm described by Conover, the probability of a Type I error can be determined exactly assuming random sampling only. For $n > 40$, an approximate critical value can be used, which is tabled by Conover when testing at the α level for $\alpha = .2, .1, .05, .02$, and $.01$.

The following example illustrates the calculations and a situation where the test might have practical value. Imagine that 10 independent studies are performed. To be concrete, suppose these 10 tests are based on Student's T. Further imagine that p values from these studies are available, but the data used to compute the p values are not. For illustrative purposes, imagine the p values are .621, .503, .203, .477, .710, .581, .329, .480, .554, and .382. So, in particular, none of the tests is significant at the .05 level. The issue here is that if we assume that for each study, the groups being compared do not differ, is it the case that Student's T provides adequate control over the probability of a Type I error? If it does, and the groups do not differ, the p values follow a uniform distribution. The ability of Student's T to control the probability of a Type I error is a serious concern because many recent papers have demonstrated that practical problems can arise, even with fairly large sample sizes. Moreover, problems with controlling the probability of a Type I error can translate into poor power when using Student's T.

Here, $F_0(x) = x$, $0 \leq x \leq 1$. For $x < 0$, $F_0(x) = 0$ and for $x > 1$, $F_0(x) = 1$. Focus on the first observation, $X_1 = .621$. The estimate of $F_1(.621)$ is $\hat{F}_1(.621) = 9/10$,

because there are nine instances where the p values are less than or equal to .621. Then $|F_1(.621) - F_0(.621)| = .9 - .621 = .279$. Repeating this for the other nine p values, we see that $T = .29$. The .05 critical value for T is .409, so we fail to reject the hypothesis that the p values follow a uniform distribution. Of course, this does not mean that the hypothesis of a uniform distribution should be accepted. Failing to reject might be because the null hypothesis is true, or because the null hypothesis is false but the power of the Kolmogorov-Smirnov test is relatively low, resulting in a Type II error.

Let $\bar{X}$ and s^2 be the usual sample mean and variance, respectively. The modification that allows one to test the hypothesis that F_1 has a normal distribution stems from Lilliefors. Letting

$$Z_i = \frac{X_i - \bar{X}}{s},$$

one merely tests the hypothesis that Z has a standard normal distribution using the test statistic previously described, only computed with the X_i values replaced by the Z_i values. The critical values differ from the case where the population mean and variances are specified, rather than estimated, but again, exact critical values can be determined.

A criticism of the one-sample version of the Kolmogorov-Smirnov test is that the Kolmogorov distance between two distributions can be small, meaning that T is relatively small, even when, in some sense, there is a substantial difference between the distributions that might have practical importance. Consider, for example, the contaminated normal distribution

$$H(x) = .9\Phi(x) + .1\Phi(x/10),$$

where Φ is the standard normal distribution. Although this contaminated normal has a relatively small Kolmogorov distance from the standard normal, its variance is 10.9 versus 1 for the standard normal. In practical terms, if the goal is to test the hypothesis that observations are sampled from a standard normal, the power of the Kolmogorov-Smirnov test to detect the true difference is relatively low. For example, if

$n = 200$, power is only about .15 when testing at the .05 level.

—Rand R. Wilcox

See also Distance; Lilliefors Test for Normality

Further Reading

- Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.

KOLMOGOROV-SMIRNOV TEST FOR TWO SAMPLES

The two-sample Kolmogorov-Smirnov test is designed to test the hypothesis that two independent groups have identical distributions. A possible appeal of the method is that it can be sensitive to differences between groups that might be routinely missed when using means, medians, or any single measure of location. For example, it might detect differences in the variances or the amount of skewness. More generally, it can detect differences between percentiles that might be missed with many alternative methods for comparing groups. Another positive feature is that it forms the basis of a graphical method for characterizing how groups differ over all the percentiles. That is, it provides an approach to assessing effect size that reveals details missed by other commonly used techniques. Moreover, the test is distribution-free, meaning that assuming random sampling only, the probability of a Type I error can be determined exactly based on the sample sizes used. Historically, the test has been described as assuming that distributions are continuous. More precisely, assuming that tied values occur with probability zero, a recursive method for

determining the exact probability of a Type I error is available. But more recently, a method that allows tied values was derived by Schroër and Trenkler.

The details are as follows. Let $X_1, \dots, X_n$ be a random sample from the first group and $Y_1, \dots, Y_m$ be a random sample from the second. Let $I_{X_i \leq x} = 1$ if $X_i \leq x$, otherwise $I_{X_i \leq x} = 0$. F_1 is estimated with

$$\hat{F}_1(x) = \frac{1}{m} \sum_{i=1}^m I_{X_i \leq x},$$

the proportion of observations less than or equal to x , and F_2 is estimated in a similar manner. The null hypothesis is

$$H_0 : F_1(x) = F_2(x), \text{ all } x$$

versus

$$H_1 : F_1(x) \neq F_2(x), \text{ for at least one } x.$$

The test statistic is based on what is sometimes called the Kolmogorov distance, which is just the maximum absolute difference between the two distributions under consideration. For convenience, let $Z_1, \dots, Z_N$ be the pooled observations where $N = m + n$. So the first m Z values correspond to $X_1, \dots, X_m$. The test statistic is

$$T = \max |F_1(Z_i) - F_2(Z_i)|,$$

the maximum being taken over all $i = 1, \dots, N$.

A variation of the Kolmogorov-Smirnov test is sometimes suggested when there is interest in detecting differences in the tails of the distributions. Let $M = nm/N$, $\lambda = n/N$, and

$$\hat{H}(x) = \lambda \hat{F}_1(x) + (1-\lambda) \hat{F}_2(x).$$

Now, the difference between any two distributions, at the value x , is estimated with

$$\frac{\sqrt{M} |\hat{F}_1(x) - \hat{F}_2(x)|}{\sqrt{\hat{H}(x)[1 - \hat{H}(x)]}}. \quad (1)$$

Then the hypothesis of identical distributions can be tested with an estimate of the largest weighted difference over all possible values of x . The test statistic is

$$D_w = \max \frac{\sqrt{M} |\hat{F}(Z_i) - \hat{G}(Z_i)|}{\sqrt{\hat{H}(Z_i)(1 - \hat{H}(Z_i))}}, \quad (2)$$

where again the maximum is taken over all values of i , $i = 1, \dots, N$, subject to $\hat{H}(Z_i)[1 - \hat{H}(Z_i)] > 0$.

Simply rejecting the hypothesis of equal distributions is not very informative. A more interesting issue is where distributions differ and by how much. A useful advance is an extension of the Kolmogorov-Smirnov test that addresses this issue. In particular, it is possible to compute confidence intervals for the difference between all of the quantiles in a manner where the probability of at least one Type I error can be determined exactly.

Suppose c is chosen so that $P(D \leq c) = 1 - \alpha$. Denote the order statistics by $X_{(1)} \leq \dots \leq X_{(n)}$ and $Y_{(1)} \leq \dots \leq Y_{(m)}$. For convenience, let $X_0 = -\infty$ and $X_{(n+1)} = \infty$. For any x satisfying $X_{(i)} \leq x < X_{(i+1)}$, let

$$k_* = \left\lceil m \left(\frac{i}{n} - \frac{c}{\sqrt{M}} \right) \right\rceil^+,$$

where $M = mn/(m + n)$ and the notation $[x]^+$ means to round up to the nearest integer. For example, $[5.1]^+ = 6$. Let

$$k^* = \left\lfloor m \left(\frac{i}{n} + \frac{c}{\sqrt{M}} \right) \right\rfloor,$$

where k^* is rounded down to the nearest integer. Then, a level $1 - \alpha$ simultaneous, distribution-free confidence band for $\Delta(x)$ ($-\infty < x < \infty$) is

$$[Y_{(k_*)} - x, Y_{(k^*+1)} - x], \quad (3)$$

where $Y_{(k_*)} = -\infty$ if $k_* < 0$ and $Y_{(k^*)} = \infty$ if $k^* \geq m + 1$. That is, with probability $1 - \alpha$, $Y_{(k_*)} - x \leq \Delta(x) < Y_{(k^*+1)} - x$ for all x . The resulting confidence band is called an S band.

Example: A portion of a study conducted by Salk dealt with weight gain in newborns who weighed at least 3,500 grams at birth. The experimental group was continuously exposed to the sound of a mother's heartbeat. The weight gains were

190 80 80 75 50 40 30 20 20 10 10 10 0
0 -10 -25 -30 -45 -60 -85

For the control group, the weight gains were

140 100 100 70 25 20 10 0 -10 -10 -25 -25
-25 -30 -30 -30 -45 -45 -45 -50 -50 -50
-60 -75 -75 -85 -85 -100 -110 -130 -130
-155 -155 -180 -240 -290

For the sake of illustration, consider computing the confidence band at $x = 77$. Because $n = 36$ and $m = 20$, $M = 12.86$. Note that the value 77 is between $Y_{(33)} = 70$ and $Y_{(34)} = 100$, so $i = 33$. From Wilcox (2005), the .05 critical value is approximately $c = .38$, so

$$k_* = \left[20 \left(\frac{33}{36} - \frac{.38}{\sqrt{12.86}} \right) \right]^+ = 17.$$

Similarly, $k^* = 20$. The 17th value in the experimental group, after putting the values in ascending order, is $X_{(17)} = 75$, $X_{(20)} = 190$, so the interval around $\Delta(77)$ is

$$(75 - 77, 190 - 77) = (-2, 113).$$

—Rand R. Wilcox

Further Reading

- Bünin, H. (2001). Kolmogorov-Smirnov and Cramer von Mises type two-sample tests with various weights. *Communications in Statistics—Theory and Methods*, 30, 847–866.
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434.
- Kim, P. J., & Jennrich, R. I. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion, D_{mn} , $m \leq n$. In H. L. Harter & D. B. Owen (Eds.), *Selected tables in mathematical statistics*, Vol. 1. Providence, RI: American Mathematical Society.

Salk, L. (1973). The role of the heartbeat in the relations between mother and infant. *Scientific American*, 235, 26–29.

Schroër, G., & Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. *Computational Statistics and Data Analysis*, 20, 185–202.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.

KR-20 AND KR-21

According to classical measurement theory, instrument reliability is the consistency that a test measures whatever it measures. A primary objective for piloting and field testing in the test construction process is to obtain an estimate of the test's reliability based on a randomly selected sample of examinees.

The reliability estimate is portable, meaning that it is a psychometric property of the test. However, it is a sound measurement practice to reassess the reliability estimate when an instrument is administered to a group that may not have been sampled randomly. It should be expected that trivial differences in the magnitude of the estimate may appear because of sampling error. This is overlooked by proponents of so-called reliability generalization, the currently popular incogitancy of conducting meta-analytic studies on reliability estimates of a test obtained in practice.

There are a variety of techniques for capturing evidence of the consistency of a test. One technique is based on internal consistency obtained within a single administration of the test. For example, consider splitting a test of N items into two parts. Let $X = N_{1, \dots, N/2}$ and $Y = N_{N/2+1, \dots, N}$. The Pearson product-moment coefficient of correlation, $r_{XY}^{\text{splithalves}}$, is called the split-halves internal consistency reliability estimate.

Internal consistency techniques have underlying assumptions. The primary assumption is that the test (or subscale of the test) is measuring a univocal factor or homogeneous construct. Statistically, internal

consistency depends on the parallel test assumption of classical measurement theory. This means that all pairs of item subsets of the test have equal means. Tests for homoscedasticity exist, but their necessity is questionable due to the practical extent that violation of this assumption inflates the magnitude of the reliability estimate, and equal intercorrelations (or correlation with the set of classical measurement theory true scores).

Another assumption of internal consistency methods is that the test is a power test rather than a speed test, which tends to inflate the magnitude of the reliability estimate. A power test is one in which sufficient time is allotted to ensure that every examinee has the opportunity to respond to every question. A speed test has a specific allotment of time to complete the test, and that amount requires quickness in order to respond to the complete set of questions. There are a variety of statistical tests to determine the “speedness” of a measurement instrument, and in its presence, there are concomitant adjustment formulas to use with internal consistency measures.

Splitting a test into two parts also assumes N is even. However, there are adjustments that permit N to be odd. These adjustments may also be used if, for some other reason, the test had to be split into unequal parts.

The most notable limitation of the split-halves approach occurs if the items in the upper half of the test represent easier material in comparison with the items appearing on the lower half of the test. This artificially depresses the reliability estimate. The problem surfaces, for example, with achievement tests because of the tendency to order items based on some logical progression, such as the sequence in which the topics are taught. One method to determine the presence of this problem is to compare the average item difficulty for the upper and the lower halves of the test.

A common index of item difficulty (P , the percentage of examinees obtaining the correct result) is obtained by $P = \frac{U + L}{N}$, where U and L are the number of examinees in the upper 27.5% and lower 27.5%, respectively, who obtained the correct answer to the test item, and $N = 27.5\% + 27.5\% = 55\%$ of the total number of examinees. The value of 27.5% probably evolved through trial and error in an effort to retain

as much information from the data set as possible, while decreasing the probability that the performance of an examinee might be classified into the wrong group. (A median split is not used to classify scores into the upper or lower group because of the high probability that a score near the median may lead to an erroneous classification.) A quick estimate useful for small samples, such as the number of students in a single classroom or the number of clients in group therapy, may be obtained by defining U and L to represent the top and bottom 10 scores, respectively, and $N = 10 + 10 = 20$.

Then, the average P value is computed for all of the items in the upper and lower halves of the test. A difference in average P value indicates that the reliability estimate will be artificially deflated. This indicates that the magnitude not only represents inconsistency of test scores due to measurement error, but also inappropriately reflects the extent to which the two parts of the test differ in terms of item difficulty.

An alternative, designed to avoid this problem, is to compute the Pearson correlation on $X = N_{1,3,5,\dots,N-1}$ and $Y = N_{2,4,6,\dots,N}$, $N = \text{even}$. In this case, $r_{XY_{\text{odd even}}}$ is called the odd-even estimate of internal consistency reliability. This will balance easy and difficult items on both X and Y . Short-cut formulas exist for $r_{XY_{\text{odd even}}}$ that require scoring only the odd or even half of the test, and the total test score.

Nevertheless, the odd-even technique may also fail if there are multiple objectives with two items per objective on the test. Consider a five-objective test with two test items per objective. The reliability estimate may be artificially depressed if, within each objective, the first (i.e., odd) item is easy and the second (i.e., even) item is difficult.

Another suggested alternative is based on a tetrad technique, where the test is split into fourths. However, there is no psychometric gain from splitting the test into more than two parts.

A compromise between split-halves and odd-even is to compute the average of the internal consistency reliability estimates obtained from both techniques. The averaged correlation will mitigate, to some extent, the impact of placement of test items, with regard to their item difficulty. Now, consider every

possible combination obtained by splitting the test into N parts of one item each. Conceptually, this is the idea behind the technique of Kuder-Richardson, which is known as the KR-20 or KR (20) because it was the 20th numbered formula in their article. The formula, in current symbols, is

$$r_{\text{KR-20}} = \frac{N}{N-1} \times \frac{\sigma_T^2 - \Sigma pq}{\sigma_T^2},$$

where N is the total number of test items, σ_T^2 is the variance of the total correct per examinee, p is the proportion of correct responses, and $q = 1 - p$. An algebraically equivalent method of determining the KR-20 is readily available and is based on a two-way analysis of variance without replication.

As an example, consider a test of $N = 6$ items taken by five examinees. The following small subset of scores was obtained during the early piloting of a subscale of the Self-Determination Knowledge Scale (SDKS). The SDKS was developed as part of a U.S. Department of Education, Office of Special Education funded initiative to develop curricula to help adolescents with and without disabilities learn to achieve their goals and become self-determined. The results are tabled below. Note that $\sigma_T^2 = 5.44$ (the computation of population variance is not shown).

Thus, the KR-20 is

$$\begin{aligned} r_{\text{KR-20}} &= \frac{6}{6-1} \times \frac{5.44 - 1.28}{5.44} \\ &= .9176 \end{aligned}$$

and because it is customary to round to two significant digits, the KR-20 is reported as .92.

Kuder and Richardson gave an approximation to the KR-20, known as KR-21, that is based on the average p and q . It simplifies the computation, given the assumption that the item difficulty is similar for all items. It obviates the need to compute indices of item

Table 1 Results for Five Examinees on a Six-Item Subscale of the SDKS

Examinee	Item						X_T (Total Correct)
	1	2	3	4	5	6	
1	1	1	1	1	1	1	6
2	1	1	1	1	1	1	6
3	1	1	1	1	0	0	4
4	1	1	0	0	0	0	2
5	0	0	0	0	0	0	0
Correct	4	4	3	3	2	2	
Incorrect	1	1	2	2	3	3	
p	.8	.8	.6	.6	.4	.4	
q	.2	.2	.4	.4	.6	.6	
pq	.16	.16	.24	.24	.24	.24	$\Sigma pq = 1.28$

Note: SDKS = Self-Determination Knowledge Scale; 0 = incorrect, 1 = correct.

difficulty. The computational formula, applied to the data in the above table, yields

$$\begin{aligned} r_{\text{KR-21}} &= \frac{N\sigma_T^2 - \bar{X}_T(N - \bar{X}_T)}{(N-1)\sigma_T^2} \\ &= \frac{(6 \times 5.44) - [3.6 \times (6 - 3.6)]}{(6-1) \times 5.44} \\ &= .8824 \end{aligned}$$

and the KR-21 would be reported as .88.

Items are sometimes assigned different weights when scoring an instrument. There are a variety of reasons for this practice (e.g., to reflect importance). Another common scoring procedure assigns different weights to a wrong response versus a missing response. An adjustment to the KR-21 to handle both of these conditions is available.

Despite the ease of computation, and the frequency that the KR-21 is reported, it assumes equal item difficulty, which, if true, would seem to obviate the need to avoid the simpler split-halves estimate. Nevertheless, the KR-21 is preferred because it is a closer approximation of the KR-20 than is the split-halves estimate of reliability.

The KR-20 is the upper bound of the KR-21. In the example from the data in the table, the KR-20 is about 4.5% larger than the KR-21. An adjustment to the KR-21 is available to close the gap, but it requires nearly as much computation as does the KR-20.

The obtained magnitude pertains to half the test length. Test length affects internal consistency reliability estimates, because as the test length decreases, the Pearson correlation decreases. The method for projecting internal consistency reliability for the complete test is known as the Spearman-Brown prophecy formula. This technique depends on the parallel test assumption of classical measurement theory and that the test is not a speed test.

The formula is $r_{SB} = \frac{2r_{IC}}{1 + 2r_{IC}}$, where r_{IC} is the value obtained via a method of internal consistency. For example, based on the KR-20 computed from the data in the table above, the estimated reliability for the entire test is

$$r_{SB} = \frac{2 \times .92}{1 + .92} = .9583$$

and would be reported as $r_{SB} = .96$.

The KR-20 is restricted to test items that are scored dichotomously (including the semantic differential, such as “agree-disagree” or “introverted-extroverted”). An extension is necessary when items are scored on a Likert scale for $N > 2$ levels (e.g., a 5-point Likert might be a scale where 1 = *strongly disagree*, 2 = *somewhat disagree*, 3 = *neutral*, 4 = *somewhat agree*, 5 = *strongly agree*). The generalization is called coefficient alpha, or Cronbach’s alpha, which is the lower-bound estimate of a Likert-scored test’s reliability.

Rules of thumb abound on what magnitude constitutes an adequate level of KR-20 or KR-21 internal consistency reliability. Obviously, as the magnitude increases, the evidence for consistency of the test increases. Commercially published tests pertaining to achievement tests, such as arithmetic or conjugation of verbs skills, are often reported to have internal consistency reliability estimates from .85 to .95 or better. Tests of less maturely developed constructs, such as giftedness, typically present much lower estimates of internal consistency reliability.

Reliability estimates are available in many statistical software packages. For example, in SPSS, select the pull-down menu sequence of Analyze | Scale | Reliability Analysis. A variety of choices, such as split-halves and Cronbach’s alpha, are readily available. Note that it is necessary to enter either correct or incorrect (i.e., 0 or 1) data into the data editor, or to write SPSS syntax to score the raw responses, prior to using the split-halves technique. This does not apply to Cronbach’s alpha if entering Likert scaled data.

—Shlomo S. Sawilowsky

Further Reading

- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Field, S., & Hoffman, A. (1994). Development of a model for self-determination. *Career Development for Exceptional Individuals*, 17, 159–169.
- Hoffman, A., Field, S., & Sawilowsky, S. (1995). *Self-Determination Knowledge Scale: Forms A & B*. Austin, TX: Pro-Ed.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Sawilowsky, S. (2002). Psychometrics versus datametrics: Comment on Vacha-Haase’s “reliability generalization” method and some EPM editorial policies. *Educational and Psychological Measurement*, 60, 157–173.
- Sawilowsky, S. (2002). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement*, 60, 196–200.
- Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, 3, 271–295.

KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE

Kruskal-Wallis analysis of variance is a statistical technique that is used to test the difference between three or more independent samples when they are of

disparate size. Remember that one of the assumptions of analysis of variance (even though the test statistic is fairly robust) is that the size of the various samples must be roughly similar and that the variances must be homogeneous.

In the case that these assumptions are not met, the Kruskal-Wallis analysis of variance technique can be used. This technique is roughly similar in use to one-way ANOVA with two or more levels of one independent variable, except that in the case of Kruskal-Wallis, this nonparametric or distribution-free statistic is very often used to test for differences between ranks.

In the following example, 30 participants are asked to rate three different types of chocolate candies: plain, peanut, and almond. The rating scale is yum, OK, and average. The hypothesis being tested is that there is a difference in ratings across the three different types of candies.

Using SPSS, the data set appears as shown below in Figure 1. As you can see, each of the 30 participants rates one of three types of candies along the three-point rating scale described above.

The procedure is performed and the output is shown in Figure 2.

Chi square is the appropriate distribution against which this test is compared for a test of the significance of the difference between the three average ranks. The output shows that there are 10 observations in each group with mean ranks ranging from 14.75 to 15.9. The chi-square value is .125, and the probability that a value of this magnitude, with 2 degrees of freedom, occurred by chance is .940. In other words, there is not a significant difference between the average ranking in each of the three groups.

—Neil J. Salkind

See also Analysis of Variance (ANOVA)

Further Reading

Smerz, J. M. (2005). Cognitive functioning in severe dementia and relationship to need driven behaviors and functional status. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 66(3B), 1737.

	Participant	Color	Taste
1	1	Peanut	Yum
2	2	Plain	OK
3	3	Almond	OK
4	4	Peanut	Average
5	5	Plain	OK
6	6	Almond	Yum
7	7	Peanut	Yum
8	8	Plain	Yum
9	9	Almond	Yum
10	10	Peanut	OK
11	11	Plain	OK
12	12	Almond	OK
13	13	Peanut	Average
14	14	Plain	Average
15	15	Almond	Average
16	16	Peanut	OK
17	17	Plain	OK
18	18	Almond	Average
19	19	Peanut	Average
20	20	Plain	Average
21	21	Almond	OK
22	22	Peanut	Yum
23	23	Plain	Yum
24	24	Almond	OK
25	25	Peanut	OK
26	26	Plain	Average
27	27	Almond	OK
28	28	Peanut	Yum
29	29	Plain	Yum
30	30	Almond	OK

Figure 1 The Data Set Where Candies Are Ranked by 30 Participants

Ranks

	Mulch color	N	Mean Rank
Taste scale	Peanut	10	14.75
	Plain	10	15.85
	Almond	10	15.90
	Total	30	

Test Statistics^{a,b}

	Taste scale
Chi-Square	.125
df	2
Asymp. Sig.	.940

a. Kruskal-Wallis Test

Figure 2 The Results of the Kruskal-Wallis Test

Woo, J., Ho, S. C., & Wong, E. M. C. (2005). Depression is the predominant factor contributing to morale as measured by the Philadelphia Geriatric Morale Scale in elderly Chinese aged 70 years and over. *International Journal of Geriatric Psychiatry*, 20(11), 1052–1059.

KUDER OCCUPATIONAL INTEREST SURVEY

The Kuder Occupational Interest Survey (KOIS; published by National Career Assessment Services) is a self-report measure of vocational interests designed to inform the educational and vocational planning and decision making of individuals in education, rehabilitation, industry, and private practice settings. The current KOIS, Form DD, evolved from continuous research and revision that began in the 1930s under the direction of Frederic Kuder.

The KOIS is appropriate for use with individuals who are high school age and older. The inventory can be administered in individual or group settings; self-administration is possible, but availability of a counselor is recommended for nuanced interpretation of scores. The items consist of 100 forced-choice triads, each requiring respondents to indicate which of three activities they prefer most and which they prefer least. The pattern of item responses is used to compute scores for scales presented in four sections of the KOIS report form: Dependability, Vocational Interest Estimates (VIEs), Occupations, and College Majors.

The first section of the KOIS report form provides a statement of the dependability of inventory results for the respondent according to indices that assess the typicality of the individual's responses. The second section of the report form presents scores for 10 VIEs, representing the following areas of vocational interests: Outdoor, Mechanical, Scientific, Computational, Persuasive, Artistic, Literary, Musical, Social Service, and Clerical. Two sets of VIE scores are provided in descending rank order, one based on female norms and one on male norms.

The third and fourth sections of the KOIS report form present scores for 109 Occupational scales and 40 College Major scales. These scores are computed

using Clemens's lambda coefficient and represent the correlation between the individual's item responses and the modal responses provided by criterion groups of satisfied women and/or men representing a specific occupation or college major. The Occupational and College Major scales also are presented in rank order (as opposed to being grouped according to some classification scheme) because the KOIS emphasizes specific information about individual occupations and majors rather than information about average relationships existing in groups. Scores within .06 lambda points of a respondent's highest score are labeled "most similar to" the interests of the criterion groups for those scales, those between .07 and .12 lambda points below the high score are labeled "next most similar," and the remaining scores are listed in order of similarity. Separate sets of scores are presented using norms for women and men.

Evidence for the reliability and validity of KOIS scale scores is generally strong and is reviewed in the KOIS general manual. In keeping with the precedent Kuder set in responding to the evolving needs of inventory users, a new instrument, the Kuder Career Search Schedule (KCSS), recently was introduced. The KCSS uses the KOIS items, but matches individual response patterns to those of a number of criterion persons (instead of criterion samples) to identify satisfied individuals whose interests are most similar to the respondent. This information is used to generate a narrative report describing the careers of the closest matching criterion persons.

—Bryan J. Dik

Further Reading

- Kuder, F. (1977). *Activity interests and occupational choice*. Chicago: Science Research Associates.
- Kuder, F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey general manual* (3rd ed.). Monterey, CA: CTB/McGraw Hill.
- Zytowski, D. G. (1992). Three generations: The continuing evolution of Frederic Kuder's interest inventories. *Journal of Counseling and Development*, 71, 245–248.

Kuder Career Planning System: <http://www.kuder.com/>

KURTOSIS

Kurtosis is commonly thought of as a measure of the “pointyness” of a frequency distribution. This is because kurtosis is the degree to which scores cluster in the tails of a frequency distribution: A platykurtic distribution has many scores in the tails (often called a heavy-tailed distribution) and so is typically quite flat, whereas a leptokurtic distribution is relatively thin in the tails and so looks quite pointy. Figure 1 shows both leptokurtic and platykurtic distributions. The leptokurtic distribution is pointier than a normal distribution; conversely, the platykurtic distribution is flatter than a normal distribution.

Kurtosis is typically measured using a scale that is centered on zero (the value of kurtosis in a normal distribution). Negative values of kurtosis represent platykurtic distributions, and positive values indicate leptokurtic distributions. If a frequency distribution has positive or negative values of kurtosis, this tells you that this distribution deviates somewhat from a normal distribution.

Values of kurtosis have associated standard errors, and these can be used to convert the value of kurtosis to a z score using the standard equation for a z score,

$$z = \frac{X - \bar{X}}{s},$$

which, if we replace the symbols with those for kurtosis, becomes

$$z_{Kurtosis} = \frac{K - \bar{K}}{SE_K}.$$

The mean value of kurtosis in the population is zero, and so the equation reduces to

$$z_{Kurtosis} = \frac{K}{SE_K}.$$

The utility of this conversion is that deviations from normality can be assessed using conventions that can be applied to any data set (regardless of the unit of

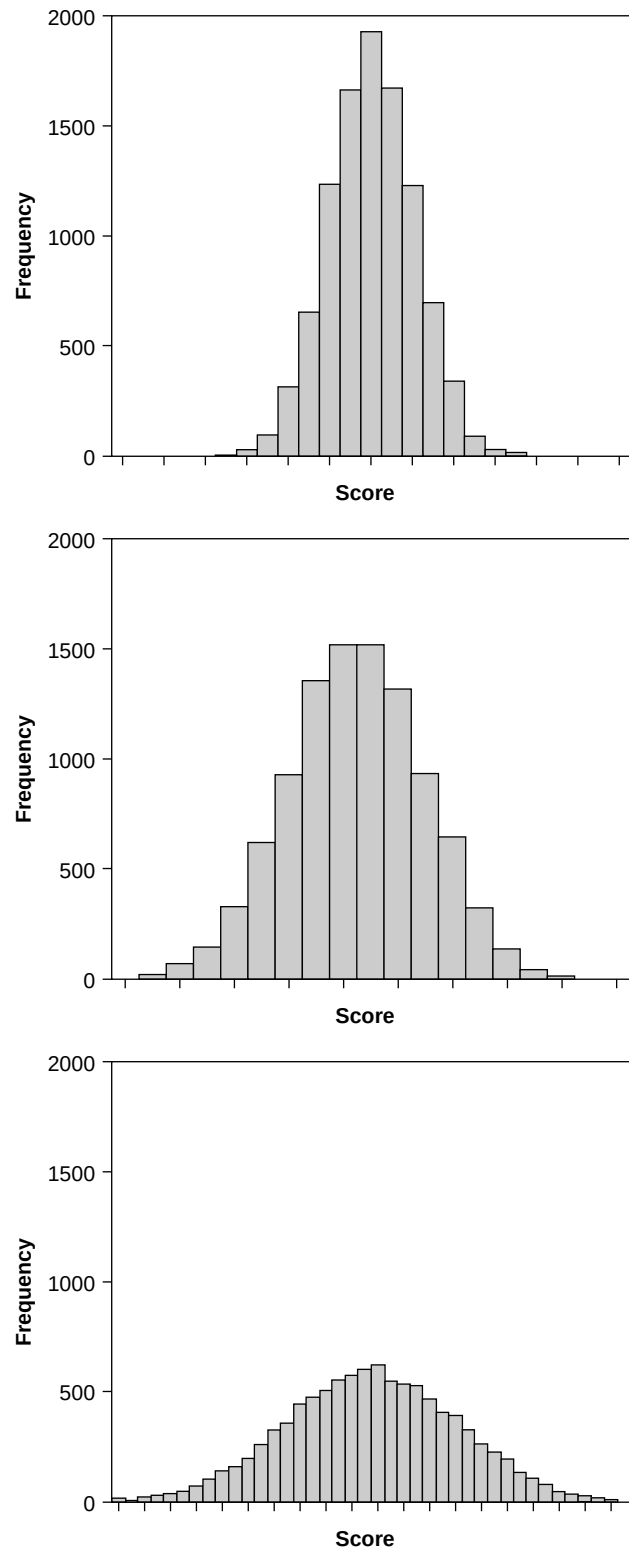


Figure 1 A Leptokurtic (top), Normal (middle), and Platykurtic (bottom) Distribution

measurement). For example, if the z associated with the value of kurtosis is greater than 1.96 (when the plus or minus sign is ignored), it is significant at $p < .05$; if it is above 2.58, then it is significant at $p < .01$; and if it is above 3.29, it is significant at $p < .001$. Although these criteria for “significant” deviations from normality can be useful, large samples will give rise to small standard errors; therefore, when sample sizes are big, significant values of kurtosis will arise from small deviations from normality. Field suggests that although these criteria can be applied to small samples, if the sample size is larger

than about 200, it is more important to look at the shape of the distribution visually (using a histogram) and to look at the value of the kurtosis statistic rather than calculating its significance.

—Andy P. Field

See also Frequency Distribution; Skewness; z Scores

Further Reading

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

L

From the errors of others, a wise man corrects his own.

—Syrus

LABORATORY BEHAVIORAL MEASURES OF IMPULSIVITY

The Laboratory Behavioral Measures of Impulsivity (available from Donald M. Dougherty, PhD, at NRLC@wfubmc.edu) are a set of four computerized measures of behavioral impulsivity: the GoStop Impulsivity Paradigm, Two Choice Impulsivity Paradigm, Single Key Impulsivity Paradigm, and Time Paradigm. This set of tests allows researchers to integrate testing of distinct types of impulsivity within a single assessment battery. The tests are widely used in research settings, and they have excellent psychometric characteristics and validity across a variety of populations (e.g., children to adults) and conditions (e.g., medication).

The GoStop Impulsivity Paradigm measures the ability to inhibit responding. Individuals taking the test are instructed to click a mouse button at a “go” stimulus (consecutive matching 5-digit numbers that appear in black against a white computer background), and to withhold responding when they see a “stop” stimulus (consecutive numbers that change in color from black to red). Responses to numbers presented with the “stop” signal reflect a failure in the ability to inhibit responding.

The Two Choice Impulsivity Paradigm (TCIP) measures the ability to forgo immediate gratification for greater rewards at a later time. Individuals taking the test must choose between shapes appearing on a computer monitor. The choices result in either a small reward (points) delivered after a short delay (e.g., 5 seconds), or a larger reward after a long delay (e.g., 15 seconds). A tendency to select more of the smaller rewards is considered impulsive, because these choices result in less optimal consequences.

The Single Key Impulsivity Paradigm (SKIP) also measures the ability to forgo immediate gratification, but does not use the forced-choice procedure of the TCIP. Instead, in the SKIP, individuals taking the test are informed that they earn rewards (points) whenever they press a button, and that these rewards are earned in proportion to how long they wait between button presses. For instance, waiting 5 seconds between presses may earn 5 points, whereas waiting 15 seconds earns 15 points. The tendency to respond more frequently for smaller rewards is considered impulsive, because these choices result in less optimal consequences.

The Time Paradigm measures the perception of the passage of time. Individuals taking the test are instructed to estimate the passage of time by pressing a button to start and stop a timer, or to hold the button continuously during the estimation interval.

A tendency to underestimate the passage of time is associated with impulsivity.

—Charles W. Mathias, Dawn M. Marsh,
and Donald M. Dougherty

Further Reading

- Dougherty, D. M., Mathias, C. W., Marsh, D. M., & Jagar, A. A. (2005). Laboratory behavioral measures of impulsivity. *Behavior Research Methods*, 37, 82–90.
- Dougherty, D. M., Mathias, C. W., & Marsh, D. M. (2003). Laboratory measures of impulsivity. In E. F. Coccaro (Ed.), *Aggression: Psychiatric assessment and treatment*. Medical Psychiatric Series No. 22 (pp. 247–265). New York: Marcel Dekker.

Laboratory Behavioral Measures of Impulsivity tasks and the laboratory where they were developed: <http://www1.wfubmc.edu/psychiatry/Research/NRLC/>

LATENT CLASS ANALYSIS

Latent class analysis (LCA) is a method or technique for identifying unlabeled groups of individuals or cases in a data set based on multivariate categorical data. LCA is used in psychology, sociology, and many other areas of application to cluster, or partition, individuals into underlying groups. The data recorded on each member of the sample are a series of measurements on categorical or qualitative variables that record discrete characteristics of the units. In many social science applications, the data consist of responses on a discrete scale to a series of questions. The latent class model (LCM) is a probability model that describes the distribution of responses in the separate groups to the several questions. Estimation of the model parameters leads to a characterization of the underlying groups in terms of likely patterns of responses to the questions.

LCA is implemented by estimating the parameters, or proportions, of the LCM. Suppose there are G latent groups in the model, and three categorical variables (A , B , and C) are measured on each case. The latent class model can be expressed as

$$p^{ABC}_{abc} = \text{sum over index } g \text{ from 1 to } G \quad (1)$$

$$p^G_g p^{AG}_{ag} p^{BG}_{bg} p^{CG}_{cg}.$$

On the left-hand side of this formula, p^{ABC}_{abc} is the probability of observing the values a , b , and c on variables A , B , and C , respectively. On the right-hand side, p^G_g is the proportion of the units in class g , p^{AG}_{ag} is the probability of a response a to Variable A among units in class g , p^{BG}_{bg} is the probability of response b to Variable B in class g , and p^{CG}_{cg} is the probability of response c to Variable C in class g . In the LCM, the Variables A , B , and C are assumed to be conditionally independent within the various groups indexed by g . This assumption is expressed in Equation 1 by the fact that the probability of response (a, b, c) in g is found by multiplying probabilities p^{AG}_{ag} , p^{BG}_{bg} , and p^{CG}_{cg} .

None of the proportions in Equation 1 needs to be known a priori in order to use LCA. If n_{abc} is the number of respondents giving response (a, b, c) to Variables (A, B, C), then the statistical likelihood for the model probabilities is a product over all patterns (a, b, c) that have been observed of p^{ABC}_{abc} raised to the n_{abc} power. The maximum likelihood estimates of the probabilities p^G_g , p^{AG}_{ag} , p^{BG}_{bg} , and p^{CG}_{cg} can be produced using numerical methods, such as the Expectation-Maximization (EM) algorithm. The term *mixture models* refers to the class of statistical models that is appropriate for populations with unlabeled subpopulations. Latent class models are mixture models that are appropriate when data are categorical and, typically, include the assumption of conditional independence.

As a hypothetical example, suppose that a group of 400 college seniors is asked to rate their level of agreement on a 3-point scale (1 = *disagree or strongly disagree*, 2 = *somewhat agree or somewhat disagree*, 3 = *agree or strongly agree*) with the following three statements: “I worry about my grades,” “I worry about whether people like me,” “I worry about money.” The largest two counts in Table 1 correspond to disagreeing with all three statements (98) and agreeing with all three statements (109).

Results of fitting a latent class model to these data are presented in Table 2.

Fifty-six percent of the observations are estimated to arise from the first class; 44% are estimated to come from the second class. Members of the first class have high probabilities of worrying about grades (80% agree or strongly agree), friends (83%), and

Table 1 Number of Responses to Three Questions by 400 College Students

<i>Hypothetical Numbers of Respondents in Each of 27 Cells of a Three-Way Table</i>				<i>Worry About Money</i>		
				<i>Disagree</i>	<i>Neutral</i>	<i>Agree</i>
Worry About Grades	Disagree	Worry	Disagree	98	19	6
		About	Neutral	24	5	3
		Friends	Agree	7	0	11
	Neutral	Worry	Disagree	18	4	2
		About	Neutral	3	1	3
		Friends	Agree	2	2	20
	Agree	Worry	Disagree	8	0	7
		About	Neutral	4	3	15
		Friends	Agree	7	19	109

Table 2 Latent Class Proportions and Probabilities

<i>Worry About . . .</i>		<i>Grades</i>	<i>Friends</i>	<i>Money</i>
Class 1 Proportion 0.56	Disagree	0.05	0.05	0.06
	Neutral	0.15	0.12	0.15
	Agree	0.80	0.83	0.79
Class 2 Proportion 0.44	Disagree	0.84	0.81	0.79
	Neutral	0.15	0.13	0.15
	Agree	0.01	0.06	0.06

money (79%). The second class has high probabilities of not worrying about grades (84% disagree or strongly disagree), friends (81%), and money (79%). These data are hypothetical and extreme, but the data and results illustrate the application of latent class models. Further analysis might relate membership in these two clusters to information on grades; participation in extracurricular activities; financial aid; family income; place of origin (small, medium, or large); family status; gender; and other factors.

It is possible to compute the probability that a respondent giving responses (a,b,c) belongs to latent class g . Using Bayes' theorem, a formula from elementary probability theory, the probability is $p_g^G p_{ag}^A p_{bg}^{BG} p_{cg}^{CG}$ divided by the value of Equation 1. Thus, the latent class analysis results can be used to cluster the response patterns and, hence, the respondents into latent classes or clusters. In practice, the number of

latent groups G is usually unknown. Some researchers approach this problem by testing the goodness-of-fit produced by latent class models with various numbers of groups. Larger models with more groups fit the observed data better, of course, but the extra groups might not be necessary to produce reasonable fit. If the groups are not needed to adequately model the data, then a smaller, more parsimonious model with fewer classes is likely a better choice. Other researchers look to the substantive interpretation of the latent classes and their probabilities to help decide on a value of G . In general, deciding on the number of classes can be a challenging problem that can be examined in various ways.

Record linkage, or exact matching, refers to the activity of linking together two or more databases on a single population. The U.S. Bureau of the Census uses record linkage in its efforts to estimate the population undercount of the decennial census.

The two files that Census links together are a sample of the decennial census and a second, independent enumeration of the population areas covered by the sample. Some individuals are enumerated in both the census and the second enumeration, whereas others are absent from one or both of the canvasses. Latent class models are used in record linkage to estimate probabilities that pairs of records, one from each of two files, correspond to a single person. The data observed for each pair of records consist of a string of zeros (indicating disagreement on a comparison of, say, last names for records in the files) and ones (indicating agreement on a comparison). The data for a particular pair of records then consist of values of several binary indicators (a type of categorical variable). Latent class analysis has been used by the U.S. Bureau of the Census to divide the pairs into groups of record pairs that are likely to be matches corresponding

to a single person and those that are likely to be nonmatches corresponding to two different people. The matches have probabilities of agreeing on comparisons that are quite high, whereas the nonmatches have lower estimated probabilities of agreement.

—Michael D. Larsen

See also Cluster Analysis; Discriminant Analysis; Mixture Models; Record Linkage

Further Reading

- Clogg, C. C., & Goodman, L. A. (1984). Latent structure-analysis of a set of multidimensional contingency-tables. *Journal of the American Statistical Association*, 79(388), 762–771.
- Goodman, L. A. (1974). Exploratory latent structure-analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215–231.
- Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Annals of Statistics*, 2(5), 911–924.
- Haberman, S. J. (1977). Product models for frequency tables involving indirect observation. *Annals of Statistics*, 5(6), 1124–1147.
- Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.
- McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

LAW OF LARGE NUMBERS

In research, a common activity is to collect a random sample of data for the purposes of estimating a parameter. As a typical example, we calculate the sample mean, or $\bar{x}$, in order to estimate the unknown mean μ of a population. Often, this is followed up by the use of some method of statistical inference; this could include the calculation of a confidence interval and/or the running of a hypothesis test. In order for our inferences to be accurate, we need some sort of assurance that our sample statistic will estimate accurately the true value of the parameter.

The Law of Large Numbers is usually described in a nonmathematically rigorous way in elementary

statistics textbooks. Essentially, if we draw a random sample from a population of any shape with a finite mean and specify how close we would like our estimate $\bar{x}$ to be to μ , eventually we can get that close as our sample size increases.

Advanced Definitions

Mathematical statisticians and probabilists define the Law of Large Numbers more precisely. First, consider the Weak Law of Large Numbers (Mood, Graybill, & Boes, 1974):

Let $f(x)$ be a probability density function with mean μ and finite variance σ^2 . Let $\bar{x}_n$ be the sample mean of a random sample of size n from $f(x)$. Choose constants α and ε such that $\varepsilon > 0$ and $0 < \alpha < 1$. If n is an integer where $n \geq \frac{\sigma^2}{\varepsilon^2 \alpha}$, then $P(-\varepsilon < \bar{x}_n - \mu < \varepsilon) \geq 1 - \alpha$.

Typically, the constant ε is chosen to be very close to zero and the theorem is used to demonstrate that as the sample size approaches infinity, the sample mean will eventually get ε -close to μ . So the formal Weak Law of Large Numbers is just the mathematical description of the Law of Large Numbers that appears in lower-level textbooks. The Strong Law of Large Numbers strengthens the Weak Law by proving that the sample means converges to μ with probability one.

Misconceptions and Application

Many people in everyday life confuse the Law of Large Numbers with a number of mistaken notions, including the so-called Law of Averages or Law of Small Numbers, where people think probabilities will even out over the short term. For instance, suppose I flip a fair coin and obtain tails five straight times. The coin is no more likely to be heads on the sixth toss, despite the Law of Averages. However, the Law of Large Numbers would tell us that the probability of obtaining heads will approach 0.50 as we flip the coin many, many times (Figure 1). In addition, Mecklin and Donnelly illustrated a novel application of the Law of Large Numbers as it applies to the Powerball lottery game (see Further Reading).

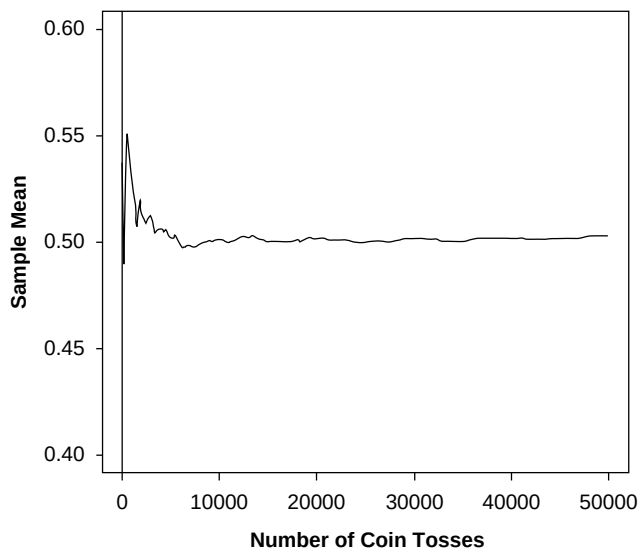


Figure 1 Coin Toss

—Christopher J. Mecklin

Further Reading

- Evans, M. J., & Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. New York: Freeman.
- Mecklin, C. J., & Donnelly, R. G. (2005). Powerball, expected value, and the law of (very) large numbers. *Journal of Statistics Education*, 13(2). Available at: <http://www.amstat.org/publications/jse/v13n2/mecklin.html>
- Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.
- Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: Freeman.

LAW SCHOOL ADMISSIONS TEST

The Law School Admissions Test (LSAT) is a half-day standardized test required for admission to all American Bar Association-approved law schools, most Canadian law schools, and many other law schools as well. The test is administered four times a year at hundreds of locations around the world. The purpose of the LSAT is to measure reasoning and logic skills thought to be essential for success in law school. These skills include the reading and comprehension of complex passages, the organization and understanding of information and the ability to draw

proper inferences from the information, the ability to think in a critical manner, and the ability to evaluate the inferences and reasoning made by others. The LSAT is among the most difficult standardized tests currently administered, with the emphasis on reasoning skills rather than knowledge.

Four sections of the test are given in a written multiple-choice format, and they have a time limit of 35 minutes each. The number of items in each section varies but is always less than 30. These four sections are used to calculate the overall score of the individual. The mean score is arbitrarily set at 150 with a standard deviation of 10, with actual scores varying from a low of 120 (−3 SD) to a high of 180 (+3 SD). These four sections consist of one reading comprehension section, one analytical reasoning section, and two different analytical reasoning sections. The test does not include any mathematical or scientific reasoning component. Due to the fact that the test norms are developed from those who actually take the test, test takers will find that their scores are generally much lower than scores on tests like the Scholastic Aptitude Test (SAT), whose normative populations are much more similar to the general population. Scores on the LSAT may vary on average as much as 1 standard deviation below scores on other standard tests, such as IQ or SAT, based on the differences in the standardization samples.

In addition, the test consists of a fifth section that is used to pretest new test items or establish the equivalence of different forms of the test. The test taker does not know which section is the experimental, unscored section. This section is not sent to law schools. Finally, the test includes a written essay question that is not scored, but which is sent to law schools along with the candidate's scores from the test. The written essay was a 30-minute, one-page test until 2005, when it was revised to a two-page, 35-minute test in which applicants receive one of two different kinds of writing prompts on a random basis.

The LSAT is administered by the Law School Admission Council (LSAC), a nonprofit organization whose members are most of the law schools in the United States and Canada. LSAT provides a number of other admission services to the member law schools, including putting together a comprehensive

packet on each applicant that includes the LSAT materials, transcripts, and letters of recommendation. Admission decisions, however, are made by each individual law school using its own criteria.

—Charles Golden

Further Reading

Law School Admission Council: <http://www.lsac.org/>

LEAST SQUARES, METHOD OF

The least squares method (LSM) is probably the most popular technique in statistics. This is due to several factors. First, most common estimators can be casted within this framework. For example, the mean of a distribution is the value that minimizes the sum of squared deviations of the scores. Second, using squares makes LSM mathematically very tractable because the Pythagorean theorem indicates that, when the error is independent of an estimated quantity, one can add the *squared* error and the *squared* estimated quantity. Third, the mathematical tools and algorithms involved in LSM (derivatives, eigendecomposition, singular value decomposition) have been well studied for a relatively long time.

LSM is one of the oldest techniques of modern statistics, and even though ancestors of LSM can be traced up to Greek mathematics, the first modern precursor is probably Galileo. The modern approach was first exposed in 1805 by the French mathematician Legendre in a now-classic memoir, but this method is somewhat older because it turned out that, after the publication of Legendre's memoir, Gauss (the famous German mathematician) contested Legendre's priority. Gauss often did not publish ideas when he thought that they could be controversial or not yet ripe, but would mention his discoveries when others published them (the way he did, for example, for the discovery of non-Euclidean geometry). And in 1809, Gauss published another memoir in which he mentioned that he had previously discovered LSM and used it as early as 1795 in estimating the orbit of an asteroid. A somewhat bitter anteriority dispute

followed (a bit reminiscent of the Leibniz-Newton controversy about the invention of calculus), which, however, did not diminish the popularity of this technique.

The use of LSM in a modern statistical framework can be traced to Galton, who used it in his work on the heritability of size, which laid the foundations of correlation and (also gave the name to) regression analysis. The two antagonistic giants of statistics, Pearson and Fisher, who did so much in the early development of statistics, used and developed LSM in different contexts (factor analysis for Pearson and experimental design for Fisher).

Today, the least squares method is widely used to find or estimate the numerical values of the parameters to fit a function to a set of data and to characterize the statistical properties of estimates. It exists with several variations: Its simpler version is called ordinary least squares (OLS), and a more sophisticated version is called weighted least squares (WLS), which often performs better than OLS because it can modulate the importance of each observation in the final solution. Recent variations of the least squares method are alternating least squares and partial least squares.

Functional Fit Example: Regression

The oldest (and still the most frequent) use of OLS was linear regression, which corresponds to the problem of finding a line (or curve) that best fits a set of data points. In the standard formulation, a set of N pairs of observations $\{Y_i, X_i\}$ is used to find a function relating the value of the dependent variable (Y) to the values of an independent variable (X). With one variable and a linear function, the prediction is given by the following equation:

$$\hat{Y} = a + bX. \quad (1)$$

This equation involves two free parameters that specify the intercept (a) and the slope (b) of the regression line. The least squares method defines the estimate of these parameters as the values that minimize the sum of the squares (hence the name

least squares) between the measurements and the model (i.e., the predicted values). This amounts to minimizing the expression,

$$\mathcal{E} = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y_i - (a + bX_i)]^2, \quad (2)$$

where $\mathcal{E}$ stands for “error,” which is the quantity to be minimized. The estimation of the parameters is obtained using basic results from calculus and, specifically, uses the property that a quadratic expression reaches its minimum value when its derivatives vanish. Taking the derivative of $\mathcal{E}$ with respect to a and b and setting them to zero gives the following set of equations (called the *normal equations*):

$$\frac{\partial \mathcal{E}}{\partial a} = 2Na + 2b \sum X_i - 2 \sum Y_i = 0 \quad (3)$$

and

$$\frac{\partial \mathcal{E}}{\partial b} = 2b \sum X_i^2 + 2a \sum X_i - 2 \sum Y_i X_i = 0. \quad (4)$$

Solving the normal equations gives the following least square estimates of a and b as

$$a = M_Y - bM_X, \quad (5)$$

where M_Y and M_X denote the means of X and Y , and

$$b = \frac{\sum (Y_i - M_Y)(X_i - M_X)}{\sum (X_i - M_X)^2}. \quad (6)$$

OLS can be extended to more than one independent variable (using matrix algebra) and to nonlinear functions.

The Geometry of Least Squares

OLS can be interpreted in a geometrical framework as an orthogonal projection of the data vector onto the space defined by the independent variable. The projection is orthogonal because the predicted values and the actual values are uncorrelated. This is illustrated in Figure 1, which depicts the case of two independent variables (vectors $\mathbf{x}_1$ and $\mathbf{x}_2$) and the data vector ($\mathbf{y}$), and shows that the error vector ($\mathbf{y} - \hat{\mathbf{y}}$) is orthogonal

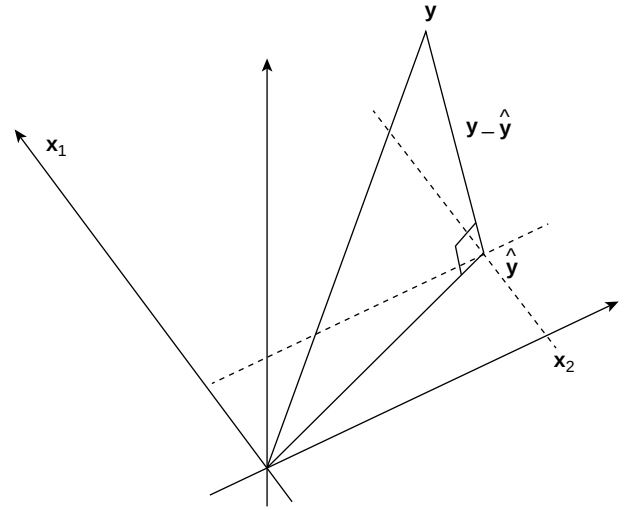


Figure 1 The Least Square Estimate of the Data Is the Orthogonal Projection of the Data Vector Onto the Independent Variable Subspace

to the least square ($\hat{\mathbf{y}}$) estimate, which lies in the subspace defined by the two independent variables.

Optimality of Least Square Estimates

OLS estimates have some strong statistical properties. Specifically, when (a) the data obtained constitute a random sample from a well-defined population, (b) the population model is linear, (c) the error has a zero expected value, (d) the independent variables are linearly independent, and (e) the error is normally distributed and uncorrelated with the independent variables (the so-called homoscedasticity assumption), then the OLS estimate is the *best linear unbiased estimate*, often denoted by the acronym “BLUE” (the five conditions and the proof are called the Gauss-Markov conditions and theorem). In addition, when the Gauss-Markov conditions hold, OLS estimates are also maximum likelihood estimates.

Weighted Least Squares

The optimality of OLS relies heavily on the homoscedasticity assumption. When the data come from different subpopulations for which an independent estimate of the error variance is available, a better estimate than OLS can be obtained using WLS, also called generalized least squares. The idea is to assign

to each observation a weight that reflects the uncertainty of the measurement. In general, the weight w_i , assigned to the i th observation, will be a function of the variance of this observation, denoted σ_i^2 . A straightforward weighting schema is to define $w_i = \sigma_i^{-1}$ (but other, more sophisticated weighted schemes can also be proposed). For the linear regression example, WLS will find the values of a and b minimizing:

$$\mathcal{E}_w = \sum_i w_i (Y_i - \hat{Y}_i)^2 = \sum_i w_i [Y_i - (a + bX_i)]^2. \quad (7)$$

Iterative Methods: Gradient Descent

When estimating the parameters of a nonlinear function with OLS or WLS, the standard approach using derivatives is not always possible. In this case, iterative methods are very often used. These methods search in a stepwise fashion for the best values of the estimate. Often, they proceed by using at each step a linear approximation of the function and refine this approximation by successive corrections. The techniques involved are known as gradient descent and Gauss-Newton approximations. They correspond to nonlinear least squares approximation in numerical analysis and nonlinear regression in statistics. Neural networks constitute a popular recent application of these techniques.

Problems With Least Squares, and Alternatives

Despite its popularity and versatility, LSM has its problems. Probably the most important drawback of LSM is its high sensitivity to outliers (i.e., extreme observations). This is a consequence of using *squares*, because squaring exaggerates the magnitude of differences (e.g., the difference between 20 and 10 is 10, but the difference between 20^2 and 10^2 is 300) and therefore gives a much stronger importance to extreme observations. This problem is addressed by using *robust* techniques, which are less sensitive to the effect of outliers. This field is currently under development and is likely to become more important in the future.

—Hervé Abdi

See also Eigendecomposition; Partial Least Square Regression; Singular and Generalized Singular Value Decomposition

Further Reading

- Abdi, H., Valentin, D., & Edelman, B. E. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
- Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Harper, H. L. (1974–1976). The method of least squares and some alternatives. Part I, II, IV, V, VI. *International Statistical Review*, 42, 147–174; 42, 235–264; 43, 1–44; 43, 125–190; 43, 269–272; 44, 113–159.
- Nocedal, J., & Wright, S. (1999). *Numerical optimization*. New York: Springer.
- Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika*, 59, 239–251.
- Seal, H. L. (1967). The historical development of the Gauss linear model. *Biometrika*, 54, 1–24.

LIFE VALUES INVENTORY

The Life Values Inventory, published by Applied Psychology Resources, contains 42 items, measures 14 relatively independent values, and takes 20 minutes to administer. The LVI was designed to help individuals clarify and prioritize their values and serve as a blueprint for decision making.

Values have long been viewed as important determinants of human behavior. Research has linked the essential aspect of values to organizational behavior, career decision making, academic performance, relationship satisfaction, and life role planning. They have also been identified as central determinants of culturally unique behavior and thus are critical to the understanding of cultural differences. The LVI was developed to provide an empirically based, easily administered values inventory with cultural sensitivity and practical utility. Applications of the LVI include career development, life role planning, adjustment and transition, retirement and leisure counseling, team building and organizational development, couples counseling, stress management, substance abuse counseling and education, and sport/performance psychology.

The 14 scales measured by the LVI are Achievement, Belonging, Concern for the Environment, Concern for Others, Creativity, Financial Prosperity, Health and Activity, Humility, Independence, Interdependence, Objective Analysis, Privacy, Responsibility, and Spirituality. The scales of the LVI were selected on the basis of a series of factor analytic studies. The scales were also reviewed for cultural sensitivity through two rounds of reviews by knowledgeable representatives of several cultural groups and subgroups.

In the process of taking the LVI, respondents are asked to both rate the strength of their values and rank them in order of importance. Respondents are first asked to rate the degree to which the beliefs contained in the 42 items are current guides to their behavior based on a 5-point Likert scale with markers for 1, 3, and 5. The markers are 1 = *seldom guides my behavior*; 3 = *sometimes guides my behavior*; and 5 = *frequently guides my behavior*. The next step in the assessment process is for individuals to rank order their most important values using the rating section as a guide. The final step is to rank the importance of the values they hope to have satisfied in each of three life roles: Work, Important Relationships, and Leisure and Community Activities. The test was updated in 2002 to improve the utility of the LVI for counseling and consulting purposes while trying to maintain the empirical validity of the 14 scales.

—R. Kelly Crace and Duane Brown

Further Reading

Brown, D., & Crace, R. K. (2002). *Facilitator's guide to the Life Values Inventory*. Williamsburg, VA: Applied Psychology Resources.

Crace, R. K., & Brown, D. (2002). *Understanding your values*. Williamsburg, VA: Applied Psychology Resources.

Life values inventory: <http://www.lifevaluesinventory.com>

LIKELIHOOD RATIO TEST

The likelihood ratio test is a test of a statistical hypothesis that uses the likelihood ratio as a test statistic. It is available in a broad class of hypothesis-testing

problems where the underlying statistical model involves a parametric family of distributions. Many well-known statistical tests are, in fact, likelihood ratio tests. In some cases, it has desirable optimality properties. For large samples, a convenient approximation is available for computing approximate p values, subject to some regularity conditions. In some less regular cases, approximate p values can be obtained via computer simulation.

Suppose that data $y = (y_1, y_2, \dots, y_n)$ are modeled as a realization of a random vector whose distribution depends on unknown parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. If the distribution is discrete, let $f(x; \theta)$ denote the probability mass function; if continuous, let $f(x; \theta)$ denote the probability density function as a function of a real vector x and the unknown parameters.

The likelihood function is

$$L(\theta; y) = f(y; \theta);$$

that is, $f(x; \theta)$ is evaluated at the data and viewed then (primarily) as a function of the unknown parameters. When the distribution is discrete, this is precisely the probability of getting the observed data as a function of the parameters; in the continuous case, the interpretation is similar, so in a certain sense, larger values of the likelihood function indicate better agreement between parameters and data.

Let Θ denote the set of all possible parameter vectors under the model, and let Θ_0 denote the set of those parameter vectors in Θ permitted under the hypothesis (this hypothesis is sometimes called the null hypothesis). Also, let Θ_1 denote the alternative, that is, the set of those parameter vectors in Θ not permitted under the hypothesis. The likelihood ratio is defined in several different but essentially equivalent ways. One is

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(\theta; y)}{\sup_{\theta \in \Theta_0} L(\theta; y)}.$$

Because larger likelihood means better agreement between parameters and data, the denominator measures the best agreement possible under the hypothesis. If the best agreement over all Θ is attained under

the hypothesis, $\Lambda = 1$. Otherwise, it is greater than 1, with better performance of the alternative over the hypothesis giving a bigger likelihood ratio. Thus, larger values of L provide more evidence against the hypothesis.

An alternate definition is

$$\Lambda_1 = \frac{\sup_{\theta \in \Theta_1} L(\theta; \mathbf{y})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{y})},$$

where the numerator measures only the best agreement under the alternative. Thus, $\Lambda = \max(\Lambda_1, 1)$, so when there is any evidence against the hypothesis, they coincide. The likelihood ratio is also sometimes defined as $1/\Lambda$ or $1/\Lambda_1$, in which case smaller values of the statistic provide more evidence against the hypothesis.

If two test statistics are increasing functions of one another, then because they always give the same p value, the tests based on them are equivalent. In many common problems, the likelihood ratio turns out to be an increasing function of a more familiar test statistic. Some examples are one-sample, two-sample, and regression t tests; analysis of variance F tests; and certain tests based on the mean of the data for binomial and Poisson models.

Power Properties

When the hypothesis and alternative each consists of a (different) single distribution, the likelihood ratio test is the most powerful test; no other test is more sensitive at detecting when the hypothesis is false. This optimality result is the essential message of the Neyman-Pearson Lemma. When the alternative and/or hypothesis consists of infinitely many points, the corresponding optimality result no longer holds in general, except in the special case when the underlying model forms a one-parameter exponential family and we are performing a one-sided test. In some other cases, the likelihood ratio test is the most powerful test among a larger group of sensible tests, but it is possible to construct examples where the likelihood ratio test has no power at all, so one should be careful in unusual situations. The book by Lehmann is a good reference for theoretical properties of the likelihood ratio test.

Calculating p Values

Once the value of the likelihood ratio statistic is obtained, the next step is to calculate the p value. In some cases, an exact p value can be computed, whereas in others, a large sample approximation is available, and in still others, computer simulation can be used to obtain an approximate p value.

If the statistic is an increasing function of another statistic with a known distribution under the hypothesis for which tail probabilities are available, we can convert to the appropriate scale and compute the p value directly. This is possible in many of the traditional tests already mentioned, where tabulated (or numerically computed) tail probabilities for distributions such as the normal, chi-squared, binomial, Poisson, F , and t distributions are readily available.

Large-Sample Behavior of $2 \log \Lambda$

In many models (those displaying so-called regularity), the maximum likelihood estimator is asymptotically normal. If so, and the hypothesis imposes m restrictions on possible values of θ , the asymptotic distribution of $2 \log \Lambda$ is a finite mixture of chi-squared distributions of the form

$$P\{2 \log \Lambda \leq x\} = \sum_{k=0}^m p_k P\{\chi_k^2 \leq x\},$$

where $P\{\chi_0^2 \leq x\} = 1\{x \geq 0\}$ is the degenerate distribution function at 0. In general, the mixing proportions p_k depend on the geometry of the model in a neighborhood of the true value θ_0 . However, under the additional condition that all points in Θ_0 are interior points of the model Θ , $p_0 = L = p_{m-1} = 0$ and $p_m = 1$, giving a pure χ_m^2 asymptotic distribution for $2 \log \Lambda$.

For the maximum likelihood estimator to be asymptotically normal, the information matrix $\mathbf{I}(\theta_0)$ must be nonsingular; here, θ_0 is the true value and $\mathbf{I}(\theta)$ has an (i, j) th element equal to

$$I_{ij}(\theta) = E \left(\frac{\dot{f}_i(\mathbf{X}; \theta)}{f(\mathbf{X}; \theta)} \frac{\dot{f}_j(\mathbf{X}; \theta)}{f(\mathbf{X}; \theta)} \right),$$

where X has distribution $f(x; \theta)$ and $f(x; \theta) = (f_1(x; \theta), \dots, f_k(x; \theta))$ is a vector of partial derivatives of $f(x; \theta)$ with respect to components of θ .

As an example where the information matrix is singular, consider the contamination model $f(x; p, \theta) = \prod_{i=1}^n [(1-p)g(x_i; 0) + pg(x_i; \theta)]$. The 2×2 information matrix has a zero on the diagonal and on the off-diagonal at either $p = 0$ or $\theta = 0$, that is, when there is no contamination and so is singular. It is known that, for certain forms of $g(x; \theta)$ in this model, under the hypothesis of no contamination $2 \log \Lambda$ diverges very slowly to infinity as the sample size grows.

Parametric Bootstrap Approximate p Values

An approximate p value can be obtained through a parametric bootstrap procedure. The idea is that for two parameter values θ and θ' that are close together, the distribution of Λ when θ is the true value should be close to the distribution of Λ when θ' is the true value. Let $p(\lambda|\theta)$ represent the probability that Λ exceeds λ when θ is the true value, let λ_{obs} denote the observed value of Λ . Then the exact p value is

$$\sup_{\theta \in \Theta_0} p(\lambda_{\text{obs}} | \theta) \approx p(\lambda_{\text{obs}} | \theta_0) \approx p(\lambda_{\text{obs}} | \hat{\theta}_0(y)),$$

because for large samples, if the hypothesis is true, the restricted (to Θ_0) maximum likelihood estimate $\hat{\theta}_0(y)$ should be close to θ_0 . The probability $p(\lambda_{\text{obs}} | \hat{\theta}_0(y))$ is the parametric bootstrap approximate p value. In general, it is difficult to obtain directly, but can itself be approximated by computer simulation.

Using the Computer

Most of the standard tests that are also likelihood ratio tests, such as t tests, F tests, and so on, can be performed on a wide variety of mathematical and statistical software systems. Even most spreadsheet programs have a facility for performing elementary statistical tests of this nature.

In nonstandard cases, one can proceed using one of the statistical systems that have a built-in programming capability, allowing users to write programs in a

simplified programming environment to perform any statistical analyses not already available. For likelihood ratio tests, this generally requires the use of optimization routines. In such cases, add-on packages are often available for nonstandard models. For example, in the R software system, the add-on library MASS has a function `fitdistr` that fits various models to univariate data by maximum likelihood. We illustrate the use of this in the example below.

Example

The twelve numbers below give time in hours between failures of the air-conditioning equipment in a Boeing 720 jet aircraft, originally cited in an article by F. Proschan.

3 5 7 18 43 85 91 98 100 130 230 487

If the failure rate is constant over time, then such times should follow an exponential distribution with density of the form $\theta e^{-\theta y}$, for $y > 0$. One way to test for constant failure rate is to embed the exponential densities in a larger model and test for exponentiality. One choice of larger model is the two-parameter family of gamma densities:

$$g(y; K, \theta) = y^{K-1} e^{-\theta y} \theta^K / \Gamma(K), \text{ for } y > 0.$$

Note that setting the shape parameter $\kappa = 1$ gives an exponential density. With κ fixed, $\log L(\kappa, \theta; \mathbf{y}) = \sum_i \log g(y_i; \kappa, \theta)$ is maximized at $\theta = \kappa / \bar{y}$, where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of the data. So if $\hat{\kappa}$ maximizes $\log L(\kappa, \kappa / \bar{y}; \mathbf{y})$, then the unrestricted maximum likelihood estimates are $(\hat{\kappa}, \hat{\kappa} / \bar{y})$, whereas the restricted estimates are $(1, 1/\bar{y})$, and so $\log \Lambda = \log L(\hat{\kappa}, \hat{\kappa} / \bar{y}; \mathbf{y}) - \log L(1, 1/\bar{y}; \mathbf{y})$.

There is no closed form for $\hat{\kappa}$; however, the maximization can be performed numerically, or we can use `fitdistr`:

```
> library(MASS)
> y <- c(3,5,7,18,43,85,91,98,100,130,230,487)
> mle <- fitdistr(y, "gamma")$estimate
> mle
      shape      rate
0.706796220 0.006537295
```

The R function *dgamma* computes the gamma density so the value of $\log \Lambda$ is given as follows:

```
> logNumer <- sum(log(dgamma(y,shape=mle[1],rate=
  mle[2]]));
> logDenom <- sum(log(dgamma(y,shape=1, rate=
  1/mean(y))));
> logLam <- (logNumer-logDenom)
> logLam
[1] 0.5494047
```

The chi-squared approximation to the p value is

```
> 1-pchisq(2*logLam,df=1)
[1] 0.2945276
```

Is the sample size large enough to trust this approximate p value? The parametric bootstrap is performed by repeating the steps above on simulated exponential samples, using the restricted maximum likelihood estimate as the rate, and computing the statistic for each simulated sample:

```
> logLamsim <- 0
> for(i in 1:1000){
+ ysim <- rexp(12,rate=1/mean(y))
+ mlesim <- fitdistr(ysim,"gamma")$estimate
+ logNumerSim <- sum(log(dgamma(ysim,shape=
  mlesim[1],rate=mlesim[2]]))
+ logDenomSim <- sum(log(dexp(ysim,rate=1/mean
  (ysim))))
+ logLamsim[i] <- (logNumerSim-logDenomSim)
+ }
> sum(logLamsim>logLam)
[1] 333
```

So, 333 of the 1,000 simulated versions of $\log \Lambda$ exceeded our observed value of 0.549, giving an approximate parametric bootstrap p value of 0.333, in reasonable agreement with the large-sample approximation of 0.295.

—Bruce Lindsay and Michael Stewart

Further Reading

- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.
- Proschan, F. (1963). Theoretical explanation of decreasing failure rate. *Technometrics*, 5, 375–383.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

LIKERT SCALING

Likert scaling (also called the *method of summated ratings*) is a multiple-item procedure for measuring attitudes. Scales resulting from the Likert method consist of a set of statements (*items*) implying favorable or unfavorable reactions to the attitudinal object. Respondents indicate their agreement to each item on a response scale most frequently ranging from 1 (representing strong disagreement) to 5 (representing strong agreement). Respondents' ratings are then summed across all items on the scale (after having reverse coded negative items), resulting in a composite score that reflects the valence and extremity of attitudes toward the object. Typical instructions and sample items are provided in Figure 1.

History and Development

Rensis Likert, in collaboration with Gardner Murphy, began development of his method of summated ratings in 1929 in response to Louis L. Thurstone's *equal-appearing intervals method*. Despite the strengths of Thurstone's approach, the method was criticized for its unwieldiness. Specifically, the method required burdensome calculations in a time before the advent of computers and required a sample of participants to act as judges in the pretesting of potential items. As a result, creating scales using the

Instructions: Please answer each item by circling the response that best reflects your level of agreement or disagreement with the statement.

1. Crime rates in this country would be significantly reduced if we had stricter gun control laws.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

2. It makes me nervous to think of how easy it is for anyone to buy a firearm in this country.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

3. Gun control legislation punishes law-abiding citizens rather than criminals.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

4. Having access to a gun would make me better able to protect my family.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

5. Given the opportunity, I would definitely purchase a firearm.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

6. It is more difficult to obtain a firearm in the United States than in most other industrialized countries.

Strongly Disagree *Disagree* *Neither Agree nor Disagree* *Agree* *Strongly Agree*

Figure 1 Sample Instructions and Items for a Likert Summated Ratings Scale

Thurstone method proved to be time-consuming. The method was also criticized by Likert and others for making a number of statistical assumptions that were unverified at the time (e.g., that the scale values of the items are independent of the attitudes of the pretesting judges). Likert developed his method with the goal of creating a reliable and valid attitude scaling procedure that was less time-consuming to construct and that did not require unnecessary assumptions. The procedure that Likert presented in 1932 has changed little throughout the years, apart from capitalizing on the statistical computation advantages offered by modern computers.

Constructing a Likert Scale

The first step in creating a Likert scale is to specify the attitudinal object. The more well-defined the subject under consideration is, the better. Once the attitudinal object has been determined, the researcher generates a pool of potential items. The aim of this phase is to create statements reflecting a broad diversity of viewpoints on the topic. Although belief or *cognitive* items (e.g., Items 1, 3, 4, and 6 in Figure 1) are most typical, feeling or *affective* items (e.g., Item 2) and behavioral tendency or *conative* items (e.g., Item 5) are also common. In contrast with some methods that use statements varying along the favorable-unfavorable continuum (e.g., Thurstone scaling), only statements that are clearly positive or negative (although not at the most extreme) should be selected for a Likert scale. The reasoning behind this recommendation is that extremely worded or neutral statements do not discriminate among respondents, that is, they do not distinguish people with very positive attitudes from people with moderately positive attitudes, negative attitudes, and so forth. Care should be taken to avoid biased or leading statements, extremely worded or neutral items, and items containing well-known facts.

The item pool usually consists of at least 25 statements, and can be as many as 50 or more. The primary reason for using multiple items rather than a single statement is that each statement may have ambiguities and subtle biases, leading people to respond in a certain way. By summing or averaging across multiple related items, the impact of biases and imperfections contained in individual items can be minimized. A secondary reason for using multiple items concerns breadth. Attitudes are often multifaceted, involving cognitions, emotions, and behavioral tendencies. A single item is unlikely to capture the full scope of the attitude in question; using multiple items potentially ameliorates this problem.

Once the researcher has generated an initial pool of items such as those illustrated in Figure 1, the third step in creating the scale is to administer the items to a sample from the population of interest.

The researcher has the respondents indicate their level of agreement to each item. Next, each item is scored for each respondent. For positive items, “strongly agree” receives five points, “agree” receives four, and so on; for negative items, the scoring is reversed. Although a 5-point response scale is most commonly employed in Likert scaling, 7-point scales are also common. Some researchers prefer to eliminate the middle, neutral category by using 4 or 6 points.

Next, each respondent’s scores are summed to obtain a total score for that individual. The highest possible score is five times the number of items, and indicates an extremely positive attitude. The lowest possible score is simply the number of items, and indicates an extremely negative attitude.

The final step in Likert’s method involves item analysis and elimination. The goal of this step is to determine which items adequately discriminate among individuals and to eliminate those that do not. Before the advent of computers, researchers commonly accomplished this task by testing whether there was a statistically significant difference between those with total scores in the top quartile and those with total scores in the bottom quartile on each particular item. With the wide availability of computers and statistical software packages, the preferred contemporary

method is to examine the correlations between respondents’ scores on individual items and their corrected total scores (i.e., the total score not including the item in question). The higher the *item-total score correlation*, the better the item is discriminating. The goal is to select items that each make a significant contribution to the variability within the total score. Items with low or no correlation with the total scores are regarded as poor items, because they contribute little to the variability in the overall score. These items do not display a strong relationship with the other items in the scale, and thus may assess a construct different from that intended by the researcher. As a result, they are discarded.

Table 1 displays hypothetical means, standard deviations, and item-total score correlations for our sample items. Examining the table, it can be seen that the mean scores for the first five items hover around the neutral point, and that the standard deviations are all in excess of one scale point, indicating a reasonable diversity of opinion among the respondents. However, the mean for the sixth item is rather low, indicating that most people strongly disagreed with this statement. In addition, the standard deviation is quite low, suggesting that the respondents in this sample did not vary greatly in their opinions. Both extreme means

Table 1 Hypothetical Means, Standard Deviations, and Discrimination Indices (r) for Gun Control Legislation Items

<i>Item</i>	<i>M</i>	<i>SD</i>	<i>r With 40-Item Scale</i>	<i>r With 20-Item Scale</i>
1. Crime rates in this country would be significantly reduced if we had stricter gun control laws.	3.55	1.24	.80	.79
2. It makes me nervous to think of how easy it is for anyone to buy a firearm in this country.	3.90	1.28	.22	—
3. Gun control legislation punishes law-abiding citizens rather than criminals.	−2.78	1.20	−.69	.67
4. Having access to a gun would make me better able to protect my family.	2.37	1.15	.90	.87
5. Given the opportunity, I would definitely purchase a firearm.	3.05	1.97	.08	—
6. It is more difficult to obtain a firearm in the United States than in most other industrialized countries.	1.34	0.42	.12	—

and low variability are warning signs that an item does not discriminate well among respondents.

The next column displays the correlations between the items and the total scores. It can be seen that both the first and fourth items correlate highly and positively with the total score, and thus discriminate well among respondents. These items contribute significantly to the overall scores and should probably be retained. The second and fifth items correlate rather weakly with the total score, suggesting that they may be measuring something different from the other items, and are therefore eliminated. The third item is interesting in that it shows a moderate to strong correlation, but it is in the wrong direction. Thus, it is actually working *against* the discrimination desired by the researcher. This may indicate an error in scoring, and if this proves to be the case, the item can be rescored and retained. In some cases, negative correlations can result from poorly worded items that may imply something different from what was intended. If respondents consistently misunderstand a statement, an extremely low or negative correlation can result. In instances of this sort, the item may be discarded or subsequently corrected and allowed to remain in the scale if the item-total correlation is as expected. The sixth item, like the second and fifth, shows a rather weak correlation. In contrast, however, this is likely due to the extremity and low variability of the responses rather than (or in addition to) the item measuring something other than what the researcher intends.

After analyzing the remaining 34 items, the researcher must decide which items to keep and which to discard. In general, the items with the highest item-total score correlations are usually kept, and these items should have reasonable variability (as indicated in our example by standard deviations). Care should be taken to maintain a balance of positive and negative items, because respondents may have a greater tendency to agree with items than to disagree (this is known as *acquiescence bias*). As well, enough items should be included in the final version of the scale to maintain adequate reliability (often measured by Cronbach's alpha). In this example, the researcher obtains a scale consisting of 10 positive and 10 negative items. The recomputed item-total score correlations in

the refined scale (as illustrated in the final column of Table 1) should be fairly similar to those in the original scale. The result of the procedure is usually a succinct, easy-to-administer, reliable, and valid attitude measure that is ready for application (and cross-validation) on other samples.

Evaluating the Likert Procedure

The reliability and validity of Likert's method has been demonstrated many times in the years since its conception. Mueller, for example, confirmed the method's validity and obtained very high interitem and test-retest reliabilities. However, the method is not without its limitations. One disadvantage of Likert scaling is that it creates a scale that is *object specific*. That is, the scale resulting from the procedure is valid only as a measure of attitudes toward the object for which it is intended. In contrast, *semantic differentials*, another type of multiple-item scaling, produce general measures that can be used with little or no alteration for a broad variety of attitudinal objects. Another criticism is that Likert's method assumes unidimensionality. It attempts to rate individuals on a single dimension of favorability; however, subsequent analyses sometimes reveal multiple dimensions. Finally, like many self-report attitude measures, the actual "neutral" point is unknown in the Likert method because this point is unlikely to correspond precisely to the midpoint of the resulting scale. This is only a problem if one is attempting to determine whether an individual respondent's score falls in an objectively favorable or unfavorable dimension; it is not of concern if one's intention is to compare the mean attitude scores of two or more groups, or to compare the mean attitude change resulting from an experimental manipulation.

Despite its limitations, the Likert procedure is generally regarded as a reliable and valid approach to measuring attitudes that is simpler to implement than the Thurstone procedure. Likert's method remains a popular and widely used method of attitude measurement.

—Leandre Fabrigar and Jay K. Wood

Further Reading

- Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
- Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In C. T. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 17–40). Thousand Oaks, CA: Sage.
- Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Orlando, FL: Harcourt Brace Jovanovich.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Mueller, D. J. (1986). *Measuring social attitudes*. New York: Teachers College Press.

Likert scales and detecting response sets (e.g., acquiescence bias): <http://www.rasch.org/erp9.htm>

LILLIEFORS TEST FOR NORMALITY

The normality assumption is at the core of a majority of standard statistical procedures, and it is important to be able to test this assumption. In addition, showing that a sample does not come from a normally distributed population is sometimes of importance per se. Among the many procedures used to test this assumption, one of the most well known is a modification of the Kolmogorov-Smirnov test of goodness of fit, generally referred to as the *Lilliefors test for normality* (or Lilliefors test, for short). This test was developed independently by Lilliefors and by Van Soest. The null hypothesis for this test is that the error is normally distributed (i.e., there is no difference between the observed distribution of the error and a normal distribution). The alternative hypothesis is that the error is not normally distributed.

Like most statistical tests, this test of normality defines a criterion and gives its sampling distribution. When the probability associated with the criterion is smaller than a given α level, the alternative hypothesis is accepted (i.e., we conclude that the sample does *not* come from a normal distribution). An interesting peculiarity of the Lilliefors test is the technique used to derive the sampling distribution of the criterion.

In general, mathematical statisticians derive the sampling distribution of the criterion using analytical techniques. However, in this case, this approach fails, and consequently, Lilliefors decided to calculate an approximation of the sampling distribution by using the Monte Carlo technique. Essentially, the procedure consists of extracting a large number of samples from a normal population and computing the value of the criterion for each of these samples. The empirical distribution of the values of the criterion gives an approximation of the sampling distribution of the criterion under the null hypothesis.

Specifically, both Lilliefors and Van Soest used, for each sample size chosen, 1,000 random samples derived from a standardized normal distribution to approximate the sampling distribution of a Kolmogorov-Smirnov criterion of goodness of fit. The critical values given by Lilliefors and Van Soest are quite similar, the relative error being of the order of 10^{-2} .

According to Lilliefors, this test of normality is more powerful than other procedures for a wide range of nonnormal conditions. Dagnelie indicated, in addition, that the critical values reported by Lilliefors can be approximated by an analytical formula. Such a formula facilitates writing computer routines because it eliminates the risk of creating errors when keying in the values of the table. Recently, Molin and Abdi refined the approximation given by Dagnelie and computed new tables using a larger number of runs (i.e., $K = 100,000$) in their simulations.

Notation

The sample for the test is made of N scores, each of them denoted X_i . The sample mean is denoted M_X and is computed as

$$M_X = \frac{1}{N} \sum_i^N X_i, \quad (1)$$

the sample variance is denoted

$$S_X^2 = \frac{\sum_i^N (X_i - M_X)^2}{N - 1}, \quad (2)$$

and the standard deviation of the sample, denoted S_X , is equal to the square root of the sample variance.

The first step of the test is to transform each of the X_i scores into z scores as follows:

$$Z_i = \frac{X_i - M_X}{S_X}. \quad (3)$$

For each Z_i score, we compute the proportion of score smaller or equal to its value: This is called the *frequency associated* with this score and it is denoted $\mathcal{S}(Z_i)$. For each Z_i score, we also compute the probability associated with this score if it comes from a “standard” normal distribution with a mean of 0 and a standard deviation of 1. We denote this probability by $\mathcal{N}(Z_i)$, and it is equal to

$$\mathcal{N}(Z_i) = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} Z_i^2 \right\}. \quad (4)$$

The criterion for the Lilliefors test is denoted L . It is calculated from the z scores, and it is equal to

$$L = \max_i \{ |\mathcal{S}(Z_i) - \mathcal{N}(Z_i)|, |\mathcal{S}(Z_i) - \mathcal{N}(Z_{i-1})| \}. \quad (5)$$

So L is the absolute value of the biggest split between the probability associated with Z_i when Z_i is normally distributed, and the frequencies actually observed. The term $|\mathcal{S}(Z_i) - \mathcal{N}(Z_{i-1})|$ is needed to take into account that, because the empirical distribution is discrete, the maximum absolute difference can occur at either endpoint of the empirical distribution.

The critical values are given by Table 2. L_{critical} is the critical value. The null hypothesis is rejected when the L criterion is greater than or equal to L_{critical} .

Numerical Example

As an illustration, we will look at an analysis of variance example for which we want to test the so-called normality assumption, which states that the within-group deviations (i.e., the residuals) are normally distributed. The data are from Abdi and correspond to memory scores obtained by 20 subjects who were assigned to one of four experimental groups (hence,

five subjects per group). The score of the s th subject in the a th group is denoted $Y_{a,s}$, and the mean of each group is denoted M_a . The within-group mean square $MS_{S(A)}$ is equal to 2.35, and it corresponds to the best estimation of the population error variance.

	<i>G. 1</i>	<i>G. 2</i>	<i>G. 3</i>	<i>G. 4</i>
	3	5	2	5
	3	9	4	4
	2	8	5	3
	4	4	4	5
	3	9	1	4
Y_a	15	35	16	21
M_a	3	7	3.2	4.2

The normality assumption states that the error is normally distributed. In the analysis of variance framework, the error corresponds to the *residuals*, which are equal to the deviations of the scores to the mean of their group. So, in order to test the normality assumption for the analysis of variance, the first step is to compute the residuals from the scores. We denote X_i the residual corresponding to the i th observation (with i going from 1 to 20). The residuals are given in the following table:

Y_{as}	3	3	2	4	3	5	9	8	4	9
X_i	0	0	-1	1	0	-2	2	1	-3	2
Y_{as}	2	4	5	4	1	5	4	3	5	4
X_i	-1.2	.8	1.8	.8	-2.2	.8	-.2	-1.2	.8	-.2

Next we transform the X_i values into Z_i values using the following formula:

$$Z_i = \frac{X_i}{\sqrt{MS_{S(A)}}}, \quad (6)$$

because $MS_{S(A)}$ is the best estimate of the population variance, and the mean of X_i is zero. Then, for each Z_i value, the frequency associated with $\mathcal{S}(Z_i)$ and the probability associated with Z_i under the normality condition $\mathcal{N}(Z_i)$ are computed [we use a table of the normal distribution to obtain $\mathcal{N}(Z_i)$]. The results are presented in Table 1.

Table 1 How to Compute the Criterion for the Lilliefors Test for Normality

X_i	N_i	F_i	Z_i	$\mathcal{S}(Z_i)$	$\mathcal{N}(Z_i)$	D_0	D_{-1}	$\max$
-3.0	1	1	-1.96	.05	.025	.025	.050	.050
-2.2	1	2	-1.44	.10	.075	.025	.075	.075
-2.0	1	3	-1.30	.15	.097	.053	.074	.074
-1.2	2	5	-.78	.25	.218	.032	.154	.154
-1.0	1	6	-.65	.30	.258	.052	.083	.083
-.2	2	8	-.13	.40	.449	.049	.143	.143
.0	3	11	.00	.55	.500	.050	.102	.102
.8	4	15	.52	.75	.699	.051	.250	.250
1.0	2	17	.65	.85	.742	.108	.151	.151
1.8	1	18	1.17	.90	.879	.021	.157	.157
2.0	2	20	1.30	1.00	.903	.097	.120	.120

Notes: N_i stands for the absolute frequency of a given value of X_i , F_i stands for the absolute frequency associated with a given value of X_i (i.e., the number of scores smaller than or equal to X_i), Z_i is the z score corresponding to X_i , $\mathcal{S}(Z_i)$ is the proportion of scores smaller than Z_i , $\mathcal{N}(Z_i)$ is the probability associated with Z_i for the standard normal distribution, $D_0 = |\mathcal{S}(Z_i) - \mathcal{N}(Z_i)|$, $D_{-1} = |\mathcal{S}(Z_i) - \mathcal{N}(Z_{i-1})|$, and $\max$ is the maximum of $\{D_0, D_{-1}\}$. The value of the criterion is $L = .250$.

The value of the criterion is

$$L = \max_i \{|\mathcal{S}(Z_i) - \mathcal{N}(Z_i)|, |\mathcal{S}(Z_i) - \mathcal{N}(Z_{i-1})|\} = .250. \quad (7)$$

Taking an α level of $\alpha = .05$, with $N = 20$, we find that the critical value is $L_{\text{critical}} = .192$ (as found in Table 2). Because L is larger than L_{critical} , the null hypothesis is rejected and we conclude that the residuals in our experiment are not distributed normally.

Numerical Approximation

The available tables for the Lilliefors test for normality typically report the critical values for a small set of alpha values. For example, the present table reports the critical values for $\alpha = [.20, .15, .10, .05, .01]$. These values correspond to the alpha values used for most tests involving only one null hypothesis, because this was the standard procedure in the late 1960s. The current statistical practice, however, favors multiple tests (maybe as a consequence of the availability of statistical packages). Because using multiple tests increases the overall Type I error (i.e., the familywise Type I

error or α_{PF}), it has become customary to recommend testing each hypothesis with a corrected α level (i.e., the Type I error per comparison, or α_{PC}) such as the Bonferroni or Šidák corrections. For example, using a Bonferroni approach with a familywise value of $\alpha_{PF} = .05$, and testing $J = 3$ hypotheses requires that each hypothesis is tested at the level of

$$\alpha_{PC} = \frac{1}{J} \alpha_{PF} = \frac{1}{3} \times .05 = .0167. \quad (8)$$

With a Šidák approach, each hypothesis will be tested at the level of

$$\begin{aligned} \alpha_{PC} &= 1 - (1 - \alpha_{PF})^{\frac{1}{J}} \\ &= 1 - (1 - .05)^{\frac{1}{3}} = .0170. \end{aligned} \quad (9)$$

As this example illustrates, both procedures are likely to require using different α levels than the ones given by the tables. In fact, it is rather unlikely that a table could be precise enough to provide the wide range of alpha values needed for multiple testing purposes. A more practical solution is to generate the critical values for any alpha value, or, alternatively, to obtain the probability associated with any value of the Kolmogorov-Smirnov criterion. Such an approach can be implemented by approximating the sampling distribution “on the fly” for each specific problem and deriving the critical values for unusual values of α .

Another approach to finding critical values for unusual values of α is to find a numerical approximation for the sampling distributions. Molin and Abdi proposed such an approximation and showed that it was accurate for at least the first two significant digits. Their procedure, somewhat complex, is better implemented with a computer and comprises two steps.

The first step is to compute a quantity, A , obtained from the following formula:

$$A = \frac{-(b_1 + N) + \sqrt{(b_1 + N)^2 - 4b_2(b_0 - L^{-2})}}{2b_2}, \quad (10)$$

Table 2 Critical Values for the Kolmogorov-Smirnov/
Lilliefors Test for Normality Obtained With
 $K = 100,000$ Samples for Each Sample Size

N	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$	N	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
4	.3027	.3216	.3456	.3754	.4129	27	.1381	.1448	.1533	.1665	.1941
5	.2893	.3027	.3188	.3427	.3959	28	.1358	.1423	.1509	.1641	.1911
6	.2694	.2816	.2982	.3245	.3728	29	.1334	.1398	.1483	.1614	.1886
7	.2521	.2641	.2802	.3041	.3504	30	.1315	.1378	.1460	.1590	.1848
8	.2387	.2502	.2649	.2875	.3331	31	.1291	.1353	.1432	.1559	.1820
9	.2273	.2382	.2522	.2744	.3162	32	.1274	.1336	.1415	.1542	.1798
10	.2171	.2273	.2410	.2616	.3037	33	.1254	.1314	.1392	.1518	.1770
11	.2080	.2179	.2306	.2506	.2905	34	.1236	.1295	.1373	.1497	.1747
12	.2004	.2101	.2228	.2426	.2812	35	.1220	.1278	.1356	.1478	.1720
13	.1932	.2025	.2147	.2337	.2714	36	.1203	.1260	.1336	.1454	.1695
14	.1869	.1959	.2077	.2257	.2627	37	.1188	.1245	.1320	.1436	.1677
15	.1811	.1899	.2016	.2196	.2545	38	.1174	.1230	.1303	.1421	.1653
16	.1758	.1843	.1956	.2128	.2477	39	.1159	.1214	.1288	.1402	.1634
17	.1711	.1794	.1902	.2071	.2408	40	.1147	.1204	.1275	.1386	.1616
18	.1666	.1747	.1852	.2018	.2345	41	.1131	.1186	.1258	.1373	.1599
19	.1624	.1700	.1803	.1965	.2285	42	.1119	.1172	.1244	.1353	.1573
20	.1589	.1666	.1764	.1920	.2226	43	.1106	.1159	.1228	.1339	.1556
21	.1553	.1629	.1726	.1881	.2190	44	.1095	.1148	.1216	.1322	.1542
22	.1517	.1592	.1690	.1840	.2141	45	.1083	.1134	.1204	.1309	.1525
23	.1484	.1555	.1650	.1798	.2090	46	.1071	.1123	.1189	.1293	.1512
24	.1458	.1527	.1619	.1766	.2053	47	.1062	.1113	.1180	.1282	.1499
25	.1429	.1498	.1589	.1726	.2010	48	.1047	.1098	.1165	.1269	.1476
26	.1406	.1472	.1562	.1699	.1985	49	.1040	.1089	.1153	.1256	.1463
						50	.1030	.1079	.1142	.1246	.1457
						>50	$\frac{0.741}{f_N}$	$\frac{0.775}{f_N}$	$\frac{0.819}{f_N}$	$\frac{0.895}{f_N}$	$\frac{1.035}{f_N}$

Notes: The intersection of a given row and column shows the critical value L_{critical} for the sample size labeling the row and the alpha level labeling the column. For $N > 50$, the critical value can be found by using $f_N = \frac{83 + N}{\sqrt{N}} - .01$.

with

$$\begin{aligned} b_2 &= 0.08861783849346 \\ b_1 &= 1.30748185078790 \\ b_0 &= 0.37872256037043. \end{aligned} \quad (11)$$

The second step implements a polynomial approximation and estimates the probability associated to a given value L as

$$\begin{aligned} \Pr(L) \approx & -.37782822932809 + 1.67819837908004A \\ & -3.02959249450445A^2 + 2.80015798142101A^3 \\ & -1.39874347510845A^4 + 0.40466213484419A^5 \\ & -0.06353440854207A^6 + 0.00287462087623A^7 \\ & +0.00069650013110A^8 - 0.00011872227037A^9 \\ & + 0.00000575586834A^{10}. \end{aligned} \quad (12)$$

For example, suppose that we have obtained a value of $L = .1030$ from a sample of size $N = 50$. (Table 2 shows that $\Pr(L) = .20$.) To estimate $\Pr(L)$, we need first to compute A , and then use this value in Equation 12. From Equation 10, we compute the estimate of A as

$$\begin{aligned} A &= \frac{-(b_1 + N) + \sqrt{(b_1 + N)^2 - 4b_2(b_0 - L^{-2})}}{2b_2} \\ &= \frac{-(b_1 + 50) + \sqrt{(b_1 + 50)^2 - 4b_2(b_0 - .1030^{-2})}}{2b_2} \\ &= 1.82402308769590. \end{aligned} \quad (13)$$

Plugging in this value of A in Equation 12 gives

$$\Pr(L) = .19840103775379 \approx .20. \quad (14)$$

As illustrated by this example, the approximated value of $\Pr(L)$ is correct for the first two decimal values.

—Hervé Abdi and Paul Molin

See also Bonferroni Test; Shapiro-Wilk Test for Normality

Further Reading

- Abdi, H. (1987). *Introduction au traitement statistique des données expérimentales*. Grenoble, France: Presses Universitaires de Grenoble.
- Dagnelie, P. (1968). A propos de l'emploi du test de Kolmogorov-Smirnov comme test de normalité. *Biométrie-Praximétrie*, 9(1), 3–13.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.
- Molin, P., & Abdi, H. (1998). *New tables and numerical approximation for the Kolmogorov-Smirnov/Lilliefors/Van Soest test of normality*. Technical report, University of Bourgogne. Available from www.utd.edu/~herve/MA_Lilliefors98.pdf
- Van Soest, J. (1967). Some experimental results concerning tests of normality. *Statistica Neerlandica*, 21, 91–97.

LINE CHART

Line charts are ideally used to demonstrate trends for data and to show how one variable is affected by another variable. Data points are connected by lines. A minimum of three points is required to make a line.

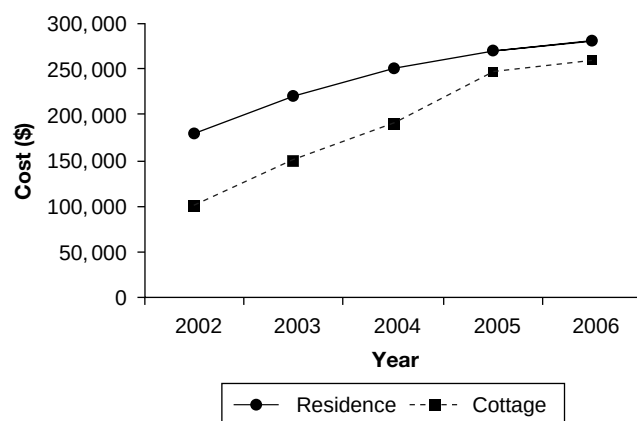


Figure 1 Cost of Housing Over 5 Years

Figure 1 is a line chart illustrating the cost of homes (principal place of residence) and cottages in the past 5 years. This graph was created using Excel. Note that even though the graph is made of discrete sets of data points, a trend is implied.

The following are recommendations for creating a line chart:

1. Gridlines help the reader estimate the value of the points.
2. The x - and y -axes are clearly labeled.
3. A legend is necessary when more than one line is presented.
4. The legend may be placed outside or inside the chart.
5. Each line is different as well as the points that make up each line in order to easily distinguish one trend from the other.
6. Too many lines (particularly if they overlap) would make the chart difficult to read.

—Adelheid A. M. Nicol

See also Bar Chart; Scattergram

Further Reading

- Bowen, R. W. (1992). *Graph it! How to make, read, and interpret graphs*. Upper Saddle River, NJ: Prentice Hall.

LINEAR REGRESSION

Linear regression is a powerful tool for testing theories about relationships among observables, and it is also useful for researchers interested in the predictive power of a set of variables. The terms *linear regression analysis* and *general linear model* are often used in the same contexts. The general linear model (GLM) is a broad class of interrelated statistical procedures focusing on linear relationships among variables or variable composites. The term *linear* is used because these techniques can be represented visually by plotting one variable against another on two-dimensional charts and using mathematical formulae for determining

where to draw one or more lines that will represent the relationships visually among the variables.

Regression analysis is the most broad, or general, form of the GLM. Hence, regression analysis forms the basis for many other statistical techniques, or, stated differently, a number of other common statistical procedures (e.g., analysis of variance, analysis of covariance, *t* test, Pearson product-moment correlation, Spearman rho correlation) are all specially designed versions of regression analysis. Furthermore, regression serves as a general framework for understanding a host of related multivariate statistics, most generally subsumed under canonical correlation analysis. An excellent nontechnical introduction to linear regression is provided by Schroeder, Sjoquist, and Stephan, and additional useful information may be found on the Web site of the Multiple Linear Regression Special Interest Group of the American Educational Research Association (<http://www.coe.unt.edu/mlrv/>).

Simple Linear Regression

Simple linear regression examines the relationship between two variables, one of which is referred to as the predictor variable (i.e., the variable that usually precedes the other), and the other of which is referred to as the criterion variable (i.e., the variable that the researcher is interested in explaining, predicting, or better understanding). Because simple regression results provide an understanding of the patterns of relationships between the two variables of interest in a given context, we often use the term *prediction* in describing the relationship. The procedure is called “simple” because it includes only one predictor variable.

As in studies employing Pearson product-moment correlation (*r*), the linear relationship between the predictor (*X*) and the criterion (*Y*) variables can be shown on a two-way scatterplot. A line of best fit drawn through the scatterplot is called the *regression line*, and the statistic representing the relationship between the two sets of points is called multiple *R*. In Pearson product-moment correlation, the coefficient *r* is used to show the strength and directionality of a relationship between two variables. A value of *r* close to zero

indicates a low or negligible correlation between two variables, whereas a value closer to $|1|$ indicates a more appreciable amount of correlation between the variables. Negative *r* values depict inverse relationships between variables, whereas positive *r* values depict direct relationships. Multiple *R* is very similar to the Pearson *r* with the exception that it is always positive in value ($0 \leq R \leq 1$) because of a set of variable weights developed as part of the analysis. Because there is only one predictor in simple linear regression, *R* will be equivalent to the absolute value of the Pearson correlation coefficient ($|r|$) between the two variables.

Predictive Equations

Predictive equations are used to determine the degree of accuracy in prediction for any given observation in the regression data set. The predictive equation in simple linear regression applies both additive (*a*) and multiplicative (*b*) weights to one variable (the predictor variable, *X*) so as to “reproduce” maximally the other variable (the criterion, or dependent variable, *Y*). The predicted *Y* score ($\hat{Y}$, or “Y-hat”) for each observation in the sample is derived from this predictive equation. The simple regression equation (“regression of *Y* on *X*”) is

$$\hat{Y} = a + bX. \quad (1)$$

This equation is the mathematical formula for a straight line. Hence, the visual resulting from the formula is called the *regression line*. Because $\hat{Y}$ is the predicted estimate of *Y*, this is often represented in the formula as follows:

$$Y \leftarrow \hat{Y} = a + bX. \quad (2)$$

Error Scores

It is important to note that the $\hat{Y}$ values are based on the *average* goodness of prediction across *Y* scores in the entire data set. Thus, some $\hat{Y}$ values will be better approximations of their corresponding *Y* than others. The accuracy of prediction for any given case in the

analysis can be found by subtracting $\hat{Y}$ from Y . The difference is the *error of prediction* (Y_e):

$$Y_e = Y - \hat{Y}. \quad (3)$$

Assumptions in Simple Linear Regression

As with any statistical procedure, simple regression analysis is based on certain assumptions about the data. As noted in various treatises on regression analysis, these assumptions include, but are not limited to, the following:

1. Simple linear regression analysis assumes that all of the variables are normally (or at least quasi-normally) distributed.
2. Regression analysis also assumes that the sample employed is randomly drawn from (or at least representative of) the population of interest.
3. Simple linear regression further assumes that a straight line will be the best way to capture the nature of the relationship between the two variables of interest. Forcing regression lines on data relationships that are curvilinear will lead to a misunderstanding about the relationship between the predictor and criterion variables.
4. Regression also is based on the assumption of “homoscedasticity” (i.e., that the conditional distribution of the Y_e scores for each value of X is an approximately normal distribution). This is also called the *constancy of error variance assumption*.

Multiple Linear Regression

Multiple linear regression is an extension of simple regression with the difference being the number of predictor variables employed. Multiple regression analyses will include at least two predictor variables ($X_1, X_2, \dots, X_k$). The linear equation for multiple regression is simply an extension of the simple regression equation:

$$Y \leftarrow \hat{Y} = a + b_1X_1 + b_2X_2 + \dots + b_kX_k. \quad (4)$$

Notice that there is a single additive (a) weight (or constant) for the equation and that each of the predictor variables (X_1 to X_k) has its own multiplicative (b) weight.

Multiple linear regression allows the researcher to look simultaneously at a host of predictor variables of interest and to examine the collective ability of these variables to predict the criterion variable. The multiple linear regression equation shown in Equation 4 above simply combines all of the predictor variables into one composite variable ($\hat{Y}$), and this composite variable then serves as a single (synthetic) predictor variable representing the host of predictors. Once the researcher determines the degree of multiple correlation (R) between the composite of predictor variables and the criterion variable, methods are then typically employed to understand further the complex set of relationships among the variables.

Holzinger and Swineford collected data on 24 tests of ability from 301 middle school students. A multiple linear regression example using scores on two of these tests (word meaning and word recognition) as predictor variables and scores on a third test (paragraph completion) as the criterion variable will be used here to illustrate the interpretation of results of a regression analysis. All output was generated using SPSS software. Linear regression results and a scatterplot depicting the relationship between Y and $\hat{Y}$ appear in Figure 1 and Figure 2, respectively.

The two predictor variables collectively account for 50.7% ($R^2 = .507$) of the variance in the criterion variable. The analysis of variance breakdown indicates that the relationship between the predictor and criterion variables (i.e., the regression sum of squares) is statistically significant at the .001 level.

Correlations, Weights, and Coefficients in Linear Regression

Linear regression yields a variety of statistical indices that aid in making interpretations of the data. We discuss here three types of indices commonly used in social science literature employing regression analyses, namely, correlations, weights, and structure coefficients.

Correlations

There are essentially two types of correlations that one might generate and interpret when conducting a regression analysis, namely, simple correlations

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.712 ^a	.507	.504	2.461

a. Predictors: (Constant), word recognition, word meaning

b. Dependent Variable: paragraph comprehension

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1854.535	2	927.267	153.139	.000 ^a
	Residual	1804.415	298	6.055		
	Total	3658.950	300			

a. predictors: (Constant), word recognition, word meaning

b. Dependent Variable: paragraph comprehension

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−1.144	2.169		−.528	.598
	Word meaning	.313	.019	.687	16.625	.000
	Word recognition	.032	.013	.104	2.526	.012

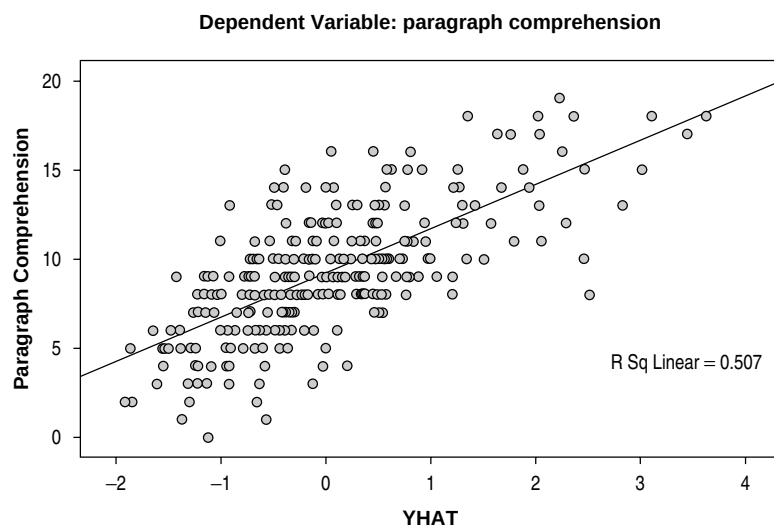
a. Dependent Variable: paragraph comprehension

Figure 1 SPSS Regression Output for Holzinger and Swineford Data

(Pearson r values) and the multiple correlation (R). Pearson r is used to express a simple linear relationship

between two variables of interest, whereas R , as previously indicated, expresses the relationship between

the composite of predictor variables and the single criterion variable. In the simple regression case, only one Pearson r can be generated, and the value of this r will be the same as the multiple R (except, possibly, for the sign of the r coefficient). In the two-predictor (multiple linear regression) case illustrated here, there are three different r values: (a) correlation between X_1 (word recognition) and Y (paragraph comprehension), (b) correlation between X_2 (word meaning) and Y , and (c) correlation between X_1 and X_2 . Correlations among these three variables and the $\hat{Y}$ generated by the analysis appear in Figure 3.

**Figure 2** Scatterplot Depicting the Relationship Between Paragraph Comprehension and Predicted Value

		Correlations			
		Word recognition	Word meaning	Paragraph comprehension	YHAT
Word recognition	Pearson Correlation Sig. (2-tailed) N	1 301	.172** .003 301	.222** .000 301	.312** .000 301
Word meaning	Pearson Correlation Sig. (2-tailed) N	.172** .003 301	1 301	.704** .000 301	.990** .000 301
Paragraph comprehension	Pearson Correlation Sig. (2-tailed) N	.222** .000 301	.704** .000 301	1 301	.712** .000 301
YHAT	Pearson Correlation Sig. (2-tailed) N	.312** .000 301	.990** .000 301	.712** .000 301	1 301

**Correlation is significant at the 0.01 level (2-tailed).

Figure 3 Correlations Among Word Recognition, Word Meaning, Paragraph Comprehension, and Predicted Value

Simple correlations between each of the predictors and the criterion variable should be reported routinely in regression studies. Based on these values, the researcher might determine whether a variable should be discarded prior to conducting the regression analysis (e.g., in the case in which the correlation with the criterion variable is nearly zero). We wish to warn, however, that this should not be simply a “fishing expedition” designed to select a set of promising variables from an array of variables for which the researcher just so happens to have data. Variables should always be selected for consideration in an analysis because they have theoretical importance, not because they are available. In the present example, word meaning is clearly the better predictor of paragraph completion.

Examining the correlations (r s) between each pair of predictor variables also can be useful. These values help the researcher to determine the degree to which the predictors are collinear with one another. Collinearity introduces a variety of problems into the regression analysis, most of which are associated with the derivation of the statistical weights for the various predictor variables. When collinearity exists, the statistical weights yielded by the regression analysis will not be uniquely determined. Furthermore, including two predictor variables in the regression analysis that are nearly perfectly collinear (i.e., that are correlated

at or near $|1.0|$) is not sensible considering that, once one of the variables is included, the second variable will offer no information that is not already provided by the first. In the present example, the correlation among the predictors is only .172, indicating that collinearity is not a major problem with this data set.

Once the Pearson r s have been consulted (and, if necessary, variables have been eliminated from the analysis), the regression analysis is then conducted. The most important correlational statistic yielded by the regression analysis is multiple R . The squared value of multiple R (R^2) represents the statistical effect size, and, as noted previously, it can be interpreted as a percent of relationship between the predictor variable set and the criterion variable. In the absence of other information, R^2 , along with the total sum of squares and the sample size, provides enough information to calculate an F statistic used in testing for statistical significance. As previously noted, the R^2 for this analysis is large and statistically significant.

Weights

Unstandardized and standardized weights yielded by the analysis appear in Figure 4. The unstandardized a and b weights yielded by the regression analysis each can be useful in interpreting the results. The multiplicative weight (b) used in the equation is also

called the *slope* and the *regression coefficient*. The additive weight (a) is also called the *Y-intercept*, the *constant*, and the *regression constant*. The a weight represents the point at which the regression line will cross the y -axis (the value of y when $x = 0$); hence, the a weight estimates the value of the dependent variable when all predictors have a value of zero. The value of the b weight indicates the number of units that the criterion value is predicted to change if the given predictor variable is increased by one unit. If the b weight is negative, the value of the predicted dependent variable will decrease by b units when x is increased by one unit. Conversely, if b is positive, the value of the predicted dependent variable will increase when x is increased. In the present example, the b weight for word meaning (i.e., .313) is appreciably larger than that for word recognition (i.e., .032).

If all of the variables in the analysis are standardized (i.e., converted to z scores), the regression equation for k variables takes the following form:

$$z_y = \beta_1 z_{x1} + \beta_2 z_{x2} + \dots + \beta_k z_{xk} + e. \quad (5)$$

Note that with standardized data, there is no longer an additive weight (e in the equation represents the error [Y_e] score, not a statistical weight). The beta (β) weights shown in Equation 5 are also known as *standardized regression coefficients* and are often used over the b weights derived using the raw data because it is easy when comparing betas to make direct comparisons as to the relative amount of weight each variable is being assigned. The b weights, on the other hand, are not as easily comparable as they are affected by the metric in which the particular predictor

variable is measured. Note that the β for word meaning is more than six times larger than that for word recognition.

Regression Structure Coefficients

As explained by Thompson and Borrello, a structure coefficient (r_s) is the correlation between each of the predictors and $\hat{Y}$. For the present example, r_s values are shown in the last column of the correlation matrix presented previously. Structure coefficients express the degree of relationship of a predictor with the predicted values of the dependent variable, or, stated differently, express the degree to which a given predictor is “reproduced” in the computation of $\hat{Y}$. In the simple regression case, the structure coefficient for the single predictor is $|1.00|$, considering that the $\hat{Y}$ is simply a linear transformation of the predictor. In multiple regression, structure coefficients are often larger for some predictors and smaller for others. In the present example, the structure coefficient for word meaning is .99, indicating that the variable is almost perfectly correlated with $\hat{Y}$. Word recognition has a much smaller structure coefficient ($r_s = .31$).

Consultation of Beta Weights and Structure Coefficients

There is a tendency for researchers to use beta weights when assessing variable contributions in linear regression. As noted earlier, these are the statistical weights applied to the standardized predictor variables in a regression. Although it seems logical to conclude that a larger beta weight implies a greater

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	−1.144	2.169		.598
	Word meaning	.313	.019	.687	.000
	Word recognition	.032	.013	.104	.012

a. Dependent Variable: paragraph comprehension

Figure 4 Unstandardized and Standardized Weights Yielded by the Regression Analysis

contribution, this is not necessarily the case. This is particularly problematic when predictor variables are collinear. In fact, in cases in which collinearity among predictors is moderate to large, beta weights are extremely subject to distortion and can lead to erroneous conclusions about the importance of variables. By contrast, regression structure coefficients are not as prone to distortion. Therefore, in determining which variables are most instrumental in regression, structure coefficients generally provide more reliable evidence than do beta weights.

—Larry Daniel, Anthony J. Onwuegbuzie,
and Nancy L. Leech

Further Reading

- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Belmont, CA: Duxbury.
- Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. Chicago: University of Chicago.
- Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.
- Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin*, 85, 410–416.
- Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage.
- Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement*, 45, 203–209.
- von Eye, A., & Schuster, C. (1998). *Regression analysis for the social sciences*. San Diego, CA: Academic Press.

Regression applet: <http://www.stat.wvu.edu/SRS/Modules/Applets/Regression/regression.html> (allows you to enter pairwise data and observe applet compute regression statistics)

Regression applet: <http://www.stattucino.com/berrie/dsl/regression/regression.html> (allows you to add one or more points to a scatterplot and observe applet draw the regression line and the correlation coefficient)

Regression applet: <http://www.mste.uiuc.edu/activity/regression/#simulation> (allows you to add up to 50 points to a scatterplot and observe applet draw the line of best fit)

LOGISTIC REGRESSION ANALYSIS

Logistic regression is a flexible tool when one variable is identified as the response (dependent) variable and it is categorical. Contingency table analysis and loglinear models can handle only categorical variables, which becomes a constraint when both categorical and continuous variables should be included in a model. Also, most of the time, researchers are interested in a particular variable as the response. More importantly, the total number of variables that can be handled by a contingency table will easily reach a limit—a table with five variables and each variable having only two values will have $2^5 = 32$ cells. Unless for a very large sample, many cells will have very few cases.

Logistic regression is better equipped than ordinary regression for modeling on a categorical response variable. The simplest categorical variable has only two different values (“dichotomous” or “binary”), such as pass or fail. A variable with multiple values without any order is multinomial, and a variable with multiple ordered values is multiordinal. Different logistic regression models should be used according to the measurement nature of the response variable. Here, the focus is on logistic regression with a binary response variable. There can be as many explanatory (independent) variables as needed, and they can be either continuous or categorical.

Several problems will arise if we still use ordinary regression analysis when the response variable is categorical. Consider the following simple linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i.$$

Note that Y can take only two possible values, say, 1 and 0, which will put a constraint on its predicted values. That is, this model will generate many predicted values that are neither 1 nor 0. And this will cause another problem when there is more than one explanatory variable. Any one of the explanatory variables may explain the majority of the very limited variation of the response variable, leaving little room

for others. In addition, the distribution of the error terms cannot be normal because the terms have only two possible values: when $Y = 1$, $\varepsilon = 1 - \alpha - \beta X$, and when $Y = 0$, $\varepsilon = -\alpha - \beta X$. Another assumption to be violated is constant error variance—the variance of X will change at different levels of X (formal expressions omitted).

These problems arise because the relationship between X and Y is not linear anymore. A transformation is needed on the categorical response variable so that it can be predicted by a linear relationship with the explanatory variables. Several transformations have been suggested, but statisticians have found the logit transformation very useful.

Odds and the Logit Transformation

In essence, logistic regression does not directly model on the value of the response variable, but on the probability that a particular value occurs. Let π be the probability that a value occurs; then $1 - \pi$ is the probability that it does not occur. The odds is the ratio of these two probabilities:

$$\text{Odds} = \frac{\pi}{1 - \pi}.$$

After taking the natural logarithm of the odds, a linear relationship between the transformed variable and the explanatory variables can be established, which is called the logistic transformation, or logit, for short:

$$\text{logit}[\pi] = \ln\left(\frac{\pi}{1 - \pi}\right).$$

The Simple Logistic Regression Model

This model has only one explanatory variable:

$$\ln\left(\frac{\pi}{1 - \pi}\right) = \alpha + \beta X.$$

The coefficient β can be interpreted in several ways. First, following the usual interpretation in ordinary

Table 1 Data for Logistic Regression

<i>Respondent</i>	<i>Support</i>	<i>Income</i>	<i>Ownership</i>
1	1	35	0
2	0	38	0
3	0	42	0
4	0	45	0
5	1	46	0
6	1	47	0
7	0	50	0
8	0	52	0
9	0	55	1
10	1	56	0
11	0	56	1
12	1	57	0
13	1	58	0
14	0	60	1
15	0	60	0
16	0	62	0
17	1	63	1
18	1	64	0
19	1	64	0
20	1	65	0
21	0	65	0
22	0	65	1
23	1	68	0
24	1	68	1
25	1	69	0
26	0	70	1
27	1	71	0
28	0	74	1
29	1	74	0
30	1	75	0
31	1	76	1
32	1	76	0
33	1	76	1
34	0	78	1
35	1	78	0
36	1	78	1
37	1	79	1
38	1	80	0
39	0	81	1
40	1	81	1
41	1	84	1
42	1	85	0
43	1	88	1
44	0	90	1
45	1	91	0
46	0	93	1
47	1	98	1
48	0	101	1
49	0	109	1
50	0	122	1

regression, we can say that with a one-unit change of X , the natural log of odds shall change by β . To many people, however, “log odds” does not make much intuitive sense.

The second interpretation comes after taking the antilog, $\frac{\pi}{1-\pi} = e^{\alpha+\beta X} = e^{\alpha}(e^{\beta})^X$. It says that with a one-unit change of X , the odds will change multiplicatively by e^{β} .

Further calculations may lead to a third interpretation. When $X = 1$, $\frac{\pi}{1-\pi} = e^{\alpha+\beta} = e^{\alpha}(e^{\beta})$. When $X = 0$, $\frac{\pi}{1-\pi} = e^{\alpha}$. The difference (or the change) of the two odds is $e^{\alpha}(e^{\beta} - 1)$. Because e^{α} is constant, $e^{\beta} - 1$ represents the change of odds. This can be written as $\% \Delta = (e^{\beta} - 1) \times 100$, where Δ is the change of the odds. With a one-unit change in X , the odds will change by $\Delta\%$.

Note that these are the interpretations when X is continuous. When it is categorical, the change of odds becomes the odds ratio. Suppose X has only two values, 0 and 1. Again, when $X = 1$, $\frac{\pi}{1-\pi} = e^{\alpha+\beta} = e^{\alpha}(e^{\beta})$, and when $X = 0$, $\frac{\pi}{1-\pi} = e^{\alpha}$. This time, the change is measured by the ratio rather than the difference between the two odds, or the odds ratio, $\frac{e^{\alpha}(e^{\beta})}{e^{\alpha}} = e^{\beta}$, which is also used in the aforementioned second interpretation. This is why computer programs like SPSS automatically produce e^{β} , but one should note that its interpretation changes according to whether X is continuous or categorical.

An Example

A survey was conducted on local residents' support for the adoption of renewable energy in a British town. Part of the results are shown in Table 1 with 50 cases and three variables.

Let π = the probability that a resident supports renewable energy, X_1 = the resident's annual household income in thousand pounds, and $X_2 = 1$ if the respondent owns a property and 0 if not.

We can have two simple logistic regression models, each with a different explanatory variable:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X_1 \text{ and } \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X_2.$$

Here are the SPSS procedures: Analyze $\rightarrow$ Regression $\rightarrow$ Binary Logistic. Bring in “support” as the dependent variable. Then, for the first model, put “income” as a covariate (because it is continuous). For now, there is no need to change other default settings, so click OK.

Table 2 SPSS Output for Simple Logistic Regression With Income

		<i>Variables in the Equation</i>					
		<i>B</i>	<i>S.E.</i>	<i>Wald</i>	<i>df</i>	<i>Sig.</i>	<i>Exp(B)</i>
Step 1(a)	Income	-.002	.016	.020	1	.886	.998
	Constant	.569	1.182	.232	1	.630	1.767

a. Variable(s) entered on step 1: income.

Table 3 SPSS Output for Simple Logistic Regression With Property Ownership

		<i>Variables in the Equation</i>					
		<i>B</i>	<i>S.E.</i>	<i>Wald</i>	<i>df</i>	<i>Sig.</i>	<i>Exp(B)</i>
Step 1(a)	Ownership(1)	1.312	.608	4.656	1	.031	3.714
	Constant	-.262	.421	.389	1	.533	.769

a. Variable(s) entered on step 1: ownership.

Table 4 SPSS Output for Multiple Logistic Regression

		<i>Variables in the Equation</i>					
		<i>B</i>	<i>S.E.</i>	<i>Wald</i>	<i>df</i>	<i>Sig.</i>	<i>Exp(B)</i>
Step 1(a)	Ownership(1)	1.788	.752	5.652	1	.017	5.978
	Income	.024	.020	1.389	1	.239	1.024
	Constant	-2.208	1.708	1.671	1	.196	.110

a. Variable(s) entered on step 1: ownership, income.

According to the output in Table 2, we have

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0.569 - 0.002 * \text{Income}.$$

The last column in the output shows the value of e^{β} , 0.998. With a one-unit change of income (1,000 pounds), the odds change by almost 1. The percentage of change (decrease) of the odds is $(0.998 - 1) * 100$, or 0.2%, showing nearly no effect of income on the probability of supporting renewable energy.

The number under “Wald” is equal to $(\hat{\beta}/ASE)^2$, where $\hat{\beta}$ is the estimated coefficient and ASE stands for asymptotic standard error, especially calculated for nonnormally distributed variables. When the sample size is large (say, more than 100), the Wald statistic has a chi-squared distribution with 1 degree of freedom. We can use it to test the null hypothesis $\beta = 0$. For this model, the Wald statistic is 0.02. Therefore, the probability of not rejecting the null hypothesis is very high (0.886). We do not have enough evidence for rejecting the hypothesis that income has no effect on the probability of supporting renewable energy.

The explanatory variable in the second model, property ownership, is categorical. In using SPSS, remember to click the “Categorical” button and put “ownership” into the “Categorical Covariates” box.

Based on Table 3, the model is

$$\ln\left(\frac{\pi}{1-\pi}\right) = -0.262 + 1.312 * \text{ownership}.$$

The last column of Table 3 shows that the odds ratio is 3.714. In other words, the odds of support for property owners is about 3.7 times of the odds of support for non-property owners. The effect is also statistically significant, with a p value of 0.031, smaller than the usual threshold of 0.05. Thus, we will reject the null hypothesis that ownership has no effect.

We can also construct a 95% confidence interval for the odds ratio. First, calculate the confidence interval for β : $(1.312 - 1.96 * 0.608, 1.312 + 1.96 * 0.608)$, or (0.12, 2.50). The confidence interval for the odds

ratio is $(e^{b-z*SE}, e^{b+z*SE}) = (e^{0.12}, e^{2.5}) = (1.13, 12.18)$. The further away the odds ratio is from 1, the more considerable the effect. Therefore, ownership has a considerable and significant effect on the probability of supporting renewable energy.

Multiple Logistic Regression Model

This is a natural extension of simple logistic regression with two or more explanatory variables. Here is a model without interaction term:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 * \text{Income} + \beta_2 * \text{ownership}.$$

SPSS output is shown in Table 4.

Because the effect of income remains insignificant after controlling that of ownership, we may want to use the simple model with ownership only. This can be confirmed by comparing the $-2 \log$ likelihood values from the two models (Table 5).

These values are an overall measurement of the difference between the observed values and the

Table 5 Model Comparison

Simple logistic regression with ownership only:

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	62.396(a)	.093	.126

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Multiple logistic regression:

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	60.950(a)	.119	.161

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

predicted values, which indicate the performance of the models. It is clear that by adding income to the simple logistic regression model with ownership only, the predicted values become closer to the observed values, but not very much (formal procedures are omitted here). Logistic regression does not have something like the R^2 in ordinary regression. The Cox & Snell R Square and the Nagelkerke R Square were thus invented to give people a sense of the model performance.

—Keming Yang

Further Reading

- Hosmer, D., & Lemeshow, S. (2001). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Pampel, F. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

SPSS regression models: http://www.spss.com/regression/data_analysis.htm (contains a downloadable spec sheet)

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Ferris, K. R., Jackofsky, E. F., & Breckenridge, B. G. (1986). Evidence for a curvilinear relationship between job performance and turnover. *Journal of Management*, 12(1), 105–111.

Regression, in general, is a technique for predicting an outcome based on a variety of variables, alone and in combination with one another. In this study, the researchers investigated the possibility of a curvilinear relationship between job performance and turnover in two diverse employee groups, using samples of 169 male accountants and 107 owner-operator truck drivers. Measures of performance for the two samples were obtained from company records. Testing for a curvilinear effect within each sample was done by regressing performance and the performance squared term on turnover, and confirming the findings by using **logistic regression**. The resulting standardized regression equations were plotted for each sample, using scores one to three standard deviations above and below the mean of the independent variables. Results indicated that performance was related to turnover in a curvilinear fashion in both professional and nonprofessional samples.

LOGLINEAR ANALYSIS

There are two general types of variables in a numeric data set, continuous (or quantitative), such as annual income, and categorical (or qualitative). Categorical variables themselves have two types, nominal, such as gender, and ordinal, such as levels of education. For a particular analysis, there may be a combination of different types of variables, which, to a large extent, determines the choice of statistical techniques. Loglinear analysis is used when (a) all the interested variables are categorical and (b) the objective is to find out which one of the interactive relationships among the variables can best explain the observed frequencies rather than explain one variable's variation with other variables. Because data with categorical variables can be presented in a contingency table, loglinear analysis is sometimes called *multiway frequency analysis* (MFA). It is also a multivariate version of chi-square analysis, dealing with variations and interactions between three or more categorical variables. Unlike in ordinary regression analysis, in loglinear analysis none of the variables is treated as a dependent variable. Rather, it is the cell count that is to be explained. Logistic regression should be used if a dependent variable is assigned and it is categorical.

Like chi-square analysis, loglinear analysis does not make any specific assumptions of distributions in the population. It only assumes that observations are independent. There is a potential problem, however, with analyzing frequencies: Especially for a small data set, some frequencies can be very small or even zero, making it very difficult to assess the reliability of results. Therefore, it is usually required that (a) there be at least five times as many cases as the number of cells in a table, and (b) when examining two variables at a time, there be an expected frequency (to be explained below) of greater than one for all cells and an expected frequency of five or more in 80% of the cells. Empty cells are usually replaced with 0.5 to avoid using the meaningless 0 as a denominator. Obviously, this is arbitrary, so a better strategy is to combine some categories in order to increase the number of cases in a particular cell.

An Example

Table 1 contains results from a survey on the students in a university. There are three categorical variables: level of study (undergraduate or postgraduate), nationality (American or international), and part-time employment during term time (employed or unemployed). The objective is to find out which model of the relationships between the variables can best predict the observed cell frequencies.

We shall use SPSS 12 to carry out the following analysis. Because the entry of cross-tabulated data is different from the entry of individual-level data, it is necessary to show all the steps here. The first is to create the variables in the “Variable View”: level of study (1 = undergraduate, 2 = postgraduate); nationality (1 = American, 0 = international); employment (1 = part-time employed, 0 = not part-time employed); and frequency. (Coding values can be different.) Then go to the “Data View” to key in the coding values and frequencies. Now you should have what is shown in Figure 1.

The last step is very important: Do *not* forget to go to Data → Weight Cases → Select “Weight case by” and bring in the variable “Frequency.” To check whether the data have been correctly entered, reproduce the table by the following procedures: Analyze → Descriptive Statistics → Crosstabs and move “Nationality” as the row variable, “Employment” as the column variable, and “Level of study” as the “Layer 1 of 1” variable. The output is shown in Table 2.

Table 2 also shows the expected values. SPSS can calculate them after you click “Cells” and select

	level	nationality	employ	frequency
1	1	1	1	15
2	1	1	0	78
3	1	0	1	8
4	1	0	0	34
5	2	1	1	26
6	2	1	0	55
7	2	0	1	9
8	2	0	0	26
9				
10				

Figure 1 Data Entry of Cross-Tabulated Data in SPSS

“Expected” below “Counts.” Below is the formula for calculating the expected value of a particular cell:

$$\text{expected frequency} = \frac{\text{column total} \times \text{row total}}{\text{overall total}}.$$

For example, the first expected frequency, 34.8, is the result of $(42 \times 112)/135$.

The Loglinear Model

The basic idea of loglinear analysis is to compare the observed frequency of a cell with its expected frequency, that is, the frequency expected to be purely by chance. Here is the logic: The bigger the difference between the two frequencies, the less likely it is that the difference arises from chance; in other words, the more likely it is that the difference is an effect of interactions between the variables. If the frequencies are purely due to chance, then there should be no or little variation across the rows or columns. Thus, the row totals and column totals at the margin (marginal totals) should be sufficient for determining the frequency in each cell. Therefore, by measuring how far away the observed frequencies are from the expected frequencies, we have a basis for discovering which variables and their interactions can best predict the observed frequencies.

Table 1 Number of Students by Level of Study, Nationality, and Employment

Level of Study	Nationality	Employment Status	
		Part-Time Employed	Not Part-Time Employed
Undergraduate	American	15	78
	International	8	34
Postgraduate	American	26	55
	International	9	26

Table 2 Reproduction of Example Data by SPSS

Nationality * Employment Status * Level of Study Crosstabulation				Employment Status		Total
Level of Study				Not Part-Time Employed	Part-Time Employed	
Undergraduate	Nationality	International	Count	34	8	42
			Expected Count	34.8	7.2	42.0
		American	Count	78	15	93
			Expected Count	77.2	15.8	93.0
	Total		Count	112	23	135
			Expected Count	112.0	23.0	135.0
Postgraduate	Nationality	International	Count	26	9	35
			Expected Count	24.4	10.6	35.0
		American	Count	55	26	81
			Expected Count	56.6	24.4	81.0
	Total		Count	81	35	116
			Expected Count	81.0	35.0	116.0

The effect of each variable is decomposed into the effect of its own (the main effect) and the effect of its interactions with other variables (the interaction effects). These effects are then connected in a linear fashion to predict the log transformation of the observed frequencies, hence the phrase “loglinear analysis.” Different combinations of the effects lead to different models.

For the above example,

F_{ijk} is the cell frequency,

λ^x is the effect of level of study,

λ^y is the effect of nationality,

λ^z is the effect of employment status,

λ is the baseline effect (when all other effects are zero).

Here are some of the possible models:

1. $\log F_{ijk} = \lambda + \lambda^x + \lambda^y + \lambda^z$, which includes only the main effects and is based on the theory that all variables are mutually independent.
2. $\log F_{ijk} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{xy} + \lambda^{xz} + \lambda^{yz}$, which includes both the main effects and the interaction

effects of all pairs of variables, suggesting that the association between any two variables remains the same at each level of the third variable.

3. $\log F_{ijk} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{xy} + \lambda^{xz} + \lambda^{yz} + \lambda^{xyz}$, which includes all the possible effects and is thus called the saturated model.

Note the hierarchical structure in the models. If a higher order interaction term is included, then all the lower order interactions and the main effects should be included.

Our objective is to make an informed decision on which of these models we should use for predicting the observed cell frequencies. The saturated model is of little use because it virtually says, “All effects matter.” Like any modeling process, we aim to find a model that is parsimonious, statistically rigorous, and substantively meaningful. Let’s focus on the first two criteria and leave the third to the researcher. We are searching for a model that fits two requirements: (a) It must have the least number of effects (or parameters), and (b) the difference between the observed frequencies and the expected frequencies must not be statistically significant. The difference is measured by

Table 3 SPSS Outputs of Loglinear Models for the Example Data

If Deleted Simple Effect is	DF	L.R. Chisq	Change	Prob	Iter
level*employ	1		6.062	.0138	2
national	1		38.480	.0000	2

Step 4

The best model has generating class

level*employ
national

Likelihood ratio chi square = .67890 DF = 3 P = .878

* * * * * H I E R A R C H I C A L L O G L I N E A R * * *

The final model has generating class

level*employ
national

The Iterative Proportional Fit algorithm converged at iteration 0.
The maximum difference between observed and fitted marginal totals is
.000 and the convergence criterion is .250

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
level	Undergra				
national	Internat				
employ	Not part	34.0	34.4	-.36	-.06
employ	Part-tim	8.0	7.1	.94	.36
national	American				
employ	Not part	78.0	77.6	.36	.04
employ	Part-tim	15.0	15.9	-.94	-.24
level	Postgrad				
national	Internat				
employ	Not part	26.0	24.8	1.15	.23
employ	Part-tim	9.0	10.7	-1.74	-.53
national	American				
employ	Not part	55.0	56.2	-1.15	-.15
employ	Part-tim	26.0	24.3	1.74	.35

Goodness-of-fit test statistics

Likelihood ratio chi square = .67890 DF = 3 P = .878
Pearson chi square = .67002 DF = 3 P = .880

Abbreviated Name	Extended Name
national	nationality

either the likelihood ratio chi-square (G^2) or Pearson's chi-square (X^2):

$$G^2 = 2 \sum (\text{observed}) \log \left(\frac{\text{observed}}{\text{fitted}} \right)$$

$$X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

G^2 is usually preferred because of its mathematical robustness.

As an interactive procedure, model selection usually starts with the saturated model, eliminates the highest order of interaction term, and then compares the new model with the saturated model; if the new model is still not good enough, the selection moves on to eliminate another effect, and so on (so-called step-wise backward elimination).

Here are the SPSS steps: Analyze → Loglinear → Model Selection, then bring the three categorical variables into "Factor(s)." Define the range of each variable: for level of study, key in 1 for minimum and 2 for maximum; for nationality and employment, key in 0 for minimum and 1 for maximum. Click OK to obtain the results. (Technical details of how to calculate the effects have been omitted.)

Results and Interpretations

Limited space does not allow all SPSS outputs to be shown here. For this example, there are four steps, and only the last one is presented in Table 3.

What we are looking for is the evidence of statistical significance after each effect term is removed. If the removal of a particular effect will make a statistically significant difference on the fitness of the model to the observed data, this effect will be retained in the model; otherwise, it will be removed in the next step. The process stops when all of the effects are statistically significant, which is indicated by the p values. For this example, two terms are kept in the final step—the interaction effect between study level and employment, and nationality. This is so because the p values are 0.0138 and 0.0000, respectively, meaning that the probability of the

values predicted by the model being equal to the observed data is very low if the term is removed. Therefore, our statistical analysis ends with the following model:

$$\log F_{ijk} = \lambda + \lambda^X + \lambda^Y + \lambda^Z + \lambda^{XZ}.$$

—Keming Yang

Further Reading

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York: HarperCollins.

Loglinear analysis is part of the advanced modules added to the basic procedures of SPSS: http://www.spss.com/advanced_models/data_analysis.htm

LONGITUDINAL/REPEATED MEASURES DATA

Longitudinal/repeated measures data arise in situations where we have multiple measures on a subject taken over time (growth over time) or a change under different treatment conditions. A simple example of longitudinal research design is when the research setting involves multiple follow-up measurements on a random sample of individuals, such as their achievement, performance, or attitude, over a period of time with logically spaced time points. Researchers across different disciplines have used different terms to describe the analysis of data obtained from repeatedly observing the same subjects. Some of the terms they have used are *longitudinal data analysis*, *within-subjects design*, *repeated-measures design*, *growth modeling*, *multilevel growth modeling*, or *individual change model*.

The simplest form of a repeated measures design is a *one-way repeated measures design*, *one-way repeated measures ANOVA design*, or *within-subjects repeated measures design*, when repeated measures on a dependent variable are observed either over time

(e.g., time, grade level) or under different treatment conditions (different type of medication for long-term illness). These times, grade levels, or treatment conditions serve as the repeated measures independent variables in the analysis.

More complex repeated measures designs have at least one between-subjects factor (e.g., gender, ethnicity) in addition to having repeated measures as within-subjects factors (e.g., time). These repeated designs with both within-subjects factors and between-subjects factors are called *repeated measures ANOVA with between-subjects factors designs* or *factorial repeated measures designs*. Below, a detailed description, assumptions, a hypothetical data set, and SPSS analysis and results of only the one-way repeated-measures ANOVA design are provided because of space limitations.

One-Way Repeated Measures ANOVA Design

In the one-way ANOVA design, all subjects are measured on all levels of the repeated measures independent variable (e.g., time or treatment conditions). Table 1 shows the representation of five subjects (S_1, S_2, S_3, S_4, S_5) in a one-way design with three time points (T_1, T_2, T_3).

Table 1 One-Way Repeated Measures Design

<i>Time</i>		
T_1	T_2	T_3
S_1	S_1	S_1
S_2	S_2	S_2
S_3	S_3	S_3
S_4	S_4	S_4
S_5	S_5	S_5

Source of Variance in One-Way Repeated Measures Design

In this design, total variability (SS_{Total}) in the measured dependent variable is partitioned into a part due to time (SS_{Time}), a part due to subjects (SS_{Subjects}), and

error ($SS_{\text{Time} \times \text{Subject}}$), which is an interaction between subject and time. Thus,

$$SS_{\text{Total}} = SS_{\text{Time}} + SS_{\text{Subjects}} + SS_{\text{Time} \times \text{Subject}} \quad (1)$$

The associated degrees of freedom in this design are partitioned as

$$df_{\text{Total}} = df_{\text{Time}} + df_{\text{Subjects}} + df_{\text{Time} \times \text{Subject}}, \quad (2)$$

where total degrees of freedom are the total number of measurements ($N = T \times S$) minus one. So,

$$df_{\text{Total}} = N - 1. \quad (3)$$

The degrees of freedom for time effect are the number of measurements over time and are represented as

$$df_{\text{Time}} = T - 1. \quad (4)$$

The degrees of freedom for subjects effect are the number of subjects minus one and are represented as

$$df_{\text{Subjects}} = S - 1. \quad (5)$$

The degrees of freedom for the error term are

$$df_{\text{Time} \times \text{Subject}} = (T - 1)(S - 1). \quad (6)$$

The corresponding mean squares (MS) are calculated by dividing sum of squares (SS) by their associated degrees of freedom (df). Thus,

$$MS_{\text{Time}} = SS_{\text{Time}} / df_{\text{Time}}, \quad (7)$$

$$MS_{\text{Subjects}} = SS_{\text{Subjects}} / df_{\text{Subjects}}, \quad (8)$$

$$MS_{\text{Time} \times \text{Subject}} = SS_{\text{Time} \times \text{Subject}} / df_{\text{Time} \times \text{Subject}}. \quad (9)$$

The F ratio for testing the null hypothesis that $\mu_{T1} = \mu_{T2} = \dots = \mu_{Tj}$, $j = 1, 2, \dots, T$ time points is the ratio of MS_{Time} and $MS_{\text{Time} \times \text{Subject}}$. Thus,

$$F = MS_{\text{Time}} / MS_{\text{Time} \times \text{Subject}}, \quad (10)$$

with $df = (T - 1), (T - 1)(S - 1)$.

Assumptions

Three assumptions underlie the one-way repeated measures design:

1. Measurements for each time point are normally distributed.
2. Subjects are independent.
3. The variances of the differences between each pair of levels of the repeated measures factor are equal. This assumption is called the “sphericity assumption,” and it is tested using Mauchly’s test. A significant Mauchly’s test indicates that the assumption of sphericity is not met. Violating this assumption inflates Type I error rate.

Analysis of Hypothetical Repeated Measures Data

Table 2 presents a small hypothetical data set in which five first-grade students are tested weekly for 3 weeks on their vocabulary learning in terms of the number of words they learn each week. The research question of interest is whether or not, on average, first-grade students’ vocabulary learning changes over the 3 weeks. Similarly, is there growth in first-grade students’ vocabulary learning over the 3 weeks?

The above hypothetical data values are entered in the SPSS data editor, where each line of data represents a subject and contains three columns of sequential dependent variable values (e.g., number of words learned each week) for each level of the repeated-measures independent variable (e.g., week1, week2, week3). To analyze these data, select the General

Linear Model option from the analysis tab of SPSS. Next, select the Repeated Measures submenu window and define “week” as the within-subjects factor. You can also request Descriptive Statistics, Eta-Square (estimate of the repeated measures effect size), Mauchly’s Test of Sphericity, and Power from the Options window.

Table 3 shows the means, standard deviations, and number of data points (subjects) in each level of the within-subjects factor (week).

The results of analyzing these hypothetical data indicate that the Mauchly’s test of sphericity is not significant (Mauchly’s $W = .918$, approximate chi-square value = .255, $df = 2$, $p = .880$). Thus, assuming sphericity, the partial eta-square value is .972 with observed power of 1.0. Also, because the assumption of sphericity is met, the results of the analysis for testing the null hypothesis of no differences between population means over the 3-week period ($H_0: \mu_{w1} = \mu_{w2} = \mu_{w3}$) using the ANOVA tests with Type III sum of squares (SS) and mean squares (MS) for within-subjects and between-subjects effects are summarized in Table 4. The results of testing the week variable reveal that there are significant differences in the means of the number of words learned over the 3-week period ($SS = 32.4$, $df = 2$, $MS = 16.2$, $F = 216$, and $p = .000$).

SPSS output also provides three alternatives for testing the same hypothesis (equality of the means over time) when the Mauchly’s test of sphericity assumption is violated. The first alternative is using the multivariate tests (e.g., Pillai’s Trace, Wilks’s Lambda, Hotelling’s Trace, Roy’s Largest Root). The second alternative is using the Greenhouse-Geisser, Huynh-Feldt, or lower-bound test adjustments. The third alternative is using a separate test of linear and

Table 2 Hypothetical Data Set for a One-Way Repeated Measures Design

Subjects	Time		
	Week ₁	Week ₂	Week ₃
S ₁	3	5	7
S ₂	2	4	5
S ₃	4	6	8
S ₄	5	7	9
S ₅	4	6	8

Table 3 Descriptive Statistics for the Hypothetical Data

Weeks	Mean	Standard Deviation	N
Week1	3.8	1.30	5
Week2	5.6	1.14	5
Week3	7.4	1.51	5

Table 4 Summary Table for One-Way Repeated Measures Analysis

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Week	32.4	2	16.2	216	.000
Subjects	20.3	4	5.1		
Week $\times$ Subjects	0.9	8	0.1		
Total	53.6	14			

quadratic trends. Because the data in the above example support the sphericity assumption, we do not need to use any of these three alternatives.

Limitations of the Traditional Repeated Measures Analysis

The above-described simple repeated measures design and other, more complex designs require the availability of the complete data, where every individual has all the measurements for all the time points as well as equal time intervals. A common problem in analyzing repeated measures/longitudinal data is that the complete data for all measurements taken at different time points for all individuals may not be available for different reasons, such as absentee at the time of data collection, drop out from the study, or any other reason. Missing data for some individuals or variations in time between consecutive measurements pose a great complication using traditional repeated-measures statistical methods.

Hierarchical linear modeling (HLM) and multi-level analysis are alternative methods to overcome the limitations of the traditional repeated measures analysis. The HLM approach is generally more flexible than the traditional repeated measures analysis in terms of its data requirements because the repeated measurements are viewed as nested within the individual rather than as the same fixed set of measurements for all individuals. Thus, both the repeated measurements and the timing of measurements may vary randomly within subjects and across subjects. Also, HLM is capable of handling longitudinal data sets with more than two levels. For instance, repeated measures for each individual (level-1) are nested

within classroom levels (level-2), which in turn could be nested within school levels (level-3). In addition, HLM is flexible in terms of fitting a polynomial growth model to the longitudinal data, where the relationship between time and the repeated measures over time is not linear. Furthermore, it is able to fit a piecewise linear growth model to a longitudinal data set for comparing growth rates during two or more different longitudinal data collection periods. Finally, HLM can include level-1 time varying covariates in the repeated data set. For example, if the repeated measures data collection points are grades (e.g., Grades 3, 4, 5), we can add to the level-1 model the number of days the students were absent in each grade as a predictor.

—Sema A. Kalaian and Rafa M. Kasim

Further Reading

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

LURIA NEBRASKA NEUROPSYCHOLOGICAL BATTERY

The extensive work of the Russian neuropsychologist A. R. Luria has long been recognized as a major contribution to clinical neuropsychology. However, his work was largely qualitative rather than quantitative in nature. The development of a standardized version of Luria's neuropsychological assessment procedures was an attempt to provide a version that could be used to generate both quantitative and qualitative information.

This undertaking was facilitated by the publication in 1975 of materials by Anne-Lise Christensen that were useful in developing the original standardized version of the Luria Nebraska Neuropsychological Battery (Form I) (LNNB). In addition, Luria summarized his investigative technique in an important paper

titled “The Neuropsychological Investigation of Patients With Localized Brain Lesions,” which was translated into English by Dr. Lawrence Majovski.

The resulting test consisted of 269 items in Form I or 279 items in Form II. The actual number of administered procedures for each form is more than 700, because many of the items include numerous repetitions of the specific procedures. In the LNNB, qualitative as well as quantitative observations are scored and used to interpret the results. The battery allows the user to make generally the same qualitative observations made by Luria and others, using more formal administration and scoring procedures. As with other, more qualitative methods, the test user is encouraged to administer each item flexibly and to continuously test the limits of the client’s cognitive abilities. The original scales for the LNNB were derived from Luria’s basic classification of his items into broad skill categories, such as motor, nonverbal auditory, visual-spatial, receptive language, and intellectual.

Form II of the battery was designed with several goals in mind. The first goal was to establish a parallel version of the battery that could be used for test-retest purposes. The second goal was to make available a set of stimulus cards that would have several advantages over the cards used in Form I. These advantages were (a) a larger size for clients with possible peripheral sensory deficits; (b) an organization that allows the examiner to use the cards more efficiently, thus reducing administration time; (c) reduction of the number of loose cards so as to further ease administration problems; (d) correction of some items that were either too difficult or too easy; (e) some changes in administration and scoring of the battery that increase the ability of the items to specify deficits

within particular areas; and (f) modifications in some items and stimuli to make the test material more familiar to users in the United States and Canada.

Since its introduction, the LNNB has generated several hundred research studies. It has been the subject of controversy from both qualitative theorists, who see the battery as too fixed and rigid, and quantitative theorists, who see the LNNB as too flexible and criticize the variation in item content within the scales of the test. Despite this, the battery is one of the most used standardized test batteries in neuropsychology and is especially useful for bedside evaluations and evaluations of clients with focal injuries as well as those who have extensive impairments that make other standard tests impractical.

—Charles Golden

Further Reading

- Christensen, A. L. (1974). *Luria’s neuropsychological investigation*. Copenhagen: Munksgaard.
- Golden, C. J., & Freshwater, S. M. (2001). Luria-Nebraska Neuropsychological Battery. In W. I. Dorfman & M. Hersen (Eds.), *Understanding psychological assessment: Perspectives on individual differences*. New York: Kluwer Academic/Plenum.
- Golden, C. J., Freshwater, S. M., & Vayalakkara, J. (2000). The Luria-Nebraska Neuropsychological Battery. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice: A guide to test interpretation and integration*. New York: Wiley.
- Moses, J. A., Jr., Golden, C. J., Ariel, R., & Gustavson, J. L. (1983). *Interpretation of the Luria-Nebraska Neuropsychological Battery, Volume 1*. New York: Grune & Stratton.
- Alexander Luria biographies: <http://www.marxists.org/archive/luria/comments/bio.htm> and http://en.wikipedia.org/wiki/Alexander_Luria

M

The safest way to double your money is to fold it over once and put it in your pocket.

—Kin Hubbard

MALE ROLE NORMS INVENTORY

Ronald F. Levant and colleagues developed the Male Role Norms Inventory (MRNI), which measures seven theoretically derived norms of traditional masculinity ideology: Avoidance of Femininity, Fear and Hatred of Homosexuals, Self-Reliance, Aggression, Achievement/Status, Non-Relational Attitudes Toward Sex, and Restrictive Emotionality. It also includes a Non-Traditional Attitudes subscale.

The MRNI consists of 57 normative statements to which subjects indicate their degree of agreement/disagreement on 7-point Likert-type scales. Examples of MRNI items: “A man should do whatever it takes to be admired and respected.” “A boy should be allowed to quit a game if he is losing.”

Scores are obtained by adding up the raw scores on individual items for each subscale and then dividing by the number of items for that subscale. For each traditional subscale, the range is 1–7, with higher scores indicating greater endorsement of traditional masculinity ideology. To obtain the Total Traditional score, add up raw scores on the six traditional subscales (i.e., excluding the Non-Traditional Attitudes subscale) and divide by 45. For the Non-Traditional Attitudes subscale, the range is also 1–7, but higher

scores indicate greater endorsement of nontraditional masculinity ideology.

The Cronbach alpha coefficients of the subscales of the original MRNI and the newly developed MRNI-R are, respectively, Avoidance of Femininity (.77, .85); Fear and Hatred of Homosexuals (.54, .91); Self-Reliance (.54, .78); Aggression (.52, .80); Achievement/Status (.67, .84.); Non-Relational Attitudes Toward Sex (.69, .79); Restrictive Emotionality (.75, .86); Non-Traditional Attitudes Toward Masculinity (.57, not used in MRNI-R); and Total Traditional scale (.84, .96).

The test-retest reliability of the MRNI (Total Traditional scale) over a 3-month time period for men was .65, and for women, .72.

Discriminant construct validity was assessed by examining the correlation of the MRNI Total Traditional scale with a theoretically distinct measure of gender—the short form of the Personal Attributes Scale (PAQ). We hypothesized that the MRNI would not be significantly correlated with PAQ and found that the MRNI Total Traditional scale was not related to the PAQ in a college student sample (for men, $r = .06$ with M, or the Masculinity scale; for females, $r = .08$ with F, or the Femininity scale). Convergent construct validity was assessed by examining the correlation of the MRNI Total Traditional scale with two

theoretically related measures of gender. We hypothesized that the MRNI would be correlated with each of these two measures and did find significant moderate correlations between the MRNI Total Traditional scale and both the Gender Role Conflict Scale-I ($r = .52$, $p < .001$) and the Masculine Gender Role Stress Scale ($r = .52$, $p < .001$).

—Ronald F. Levant

Further Reading

- Levant, R. F., & Fischer, J. (1998). The Male Role Norms Inventory. In C. M. Davis, W. H. Yarber, R. Bauserman, G. Schreer, & S. L. Davis (Eds.), *Sexuality-related measures: A compendium* (2nd ed., pp. 469–472). Thousand Oaks, CA: Sage.
- Levant, R. F., & Richmond, K. (n.d.). *A program of research on masculinity ideologies using the Male Role Norms Inventory*. Manuscript submitted for publication.

Dr. Levant: <http://www.DrRonaldLevant.com>

MALTHUS, THOMAS (1766–1834)

Malthus could be described as England's first academic economist. He became famous for his 1798 pamphlet "An Essay on the Principle of Population as It Affects the Future Improvement on Society, With Remarks on the Speculation of Mr. Godwin, M. Condorcet and Other Writers." His claim, sensational at the time, was that any growing population would eventually be unable to sustain itself.

Malthus was born in Dorking, Surrey, on February 13, 1766. He was the second son in a prosperous family of eight (two boys, six girls). Educated at home, he went up to Jesus College, Cambridge, in 1784, becoming a college fellow in 1793. He was ordained a minister of the Church of England in 1788, and in 1796, he became curate of Okewood Chapel near Albury in Surrey. For the next 8 years, he divided his time between Surrey (living with his parents) and Cambridge. When at home, he and his father had long debates on the

nation's economy, and they are what led to his famous pamphlet.

Publication of the pamphlet, which was expanded in subsequent editions by the inclusion of population data that confirmed his reasoning, turned Malthus into an intellectual celebrity. In the 1803 edition of his work, Malthus suggested that one way of reducing population would be to give to the poor those rights possessed by the middle class, such as universal suffrage, state-run education, the elimination of the Poor Laws, and the establishment of a national labor market. In effect, he was arguing that an increase in income would lead to a decrease in family size. These were controversial views, unpopular with many, yet acknowledged by others as having merit.

In 1804, Malthus gave up his fellowship and got married. The following year, he was appointed Professor of Modern History and Political Economy at the East India College in Haileybury (now Haileybury College), a position he held for the rest of his life. Malthus remained prominent in economic discussions for the rest of his life. In 1814 he published *Observations on the Effects of the Corn Laws, and of a Rise or Fall in the Price of Corn on the Agriculture and General Wealth of the Country*, which presented both sides of the argument for these laws. By the following year, however, it was clear that he personally was in favor of restrictions on the importation of foreign corn.

In 1820, he published *Principles of Political Economy*, in which he disagreed with the "classical approach" proposed by David Ricardo, a good friend.

Malthus died in Bath on December 29, 1834, following a Christmas visit to his in-laws.

—Graham Upton

See also Probability Sampling

Further Reading

- Jensen, A.-M., Knutsen, T., & Skonhoft, A. (Eds.). (2003). *Visiting Malthus: The man, his times, the issues*. Copenhagen: Copenhagen Business School Press.

Thomas Malthus biography with comprehensive links to other sites: <http://cepa.newschool.edu/het/profiles/malthus.htm>

MANN-WHITNEY U TEST (WILCOXON RANK-SUM TEST)

The Wilcoxon Rank-Sum Test was developed by Wilcoxon in 1945, and it is useful when comparing the location of two independent samples. A slightly different version of the test was later introduced by Mann and Whitney in 1947. Therefore, it is sometimes referred to as the Wilcoxon Mann-Whitney test.

The underlying assumptions of the Wilcoxon Rank-Sum Test are that the scores are independent and come from a continuous probability distribution. The null hypothesis is $H_0: f_i(x) = g_i(x)$, or the two samples come from identical distributions. The alternative hypotheses are $H_a: f_i(x) \neq g_i(x)$ (two-tail), $H_a: f_i(x) < g_i(x)$ (lower-tail), or $H_a: f_i(x) > g_i(x)$ (upper-tail).

The null hypothesis suggests that the Wilcoxon Rank-Sum is a test of general differences. Even though the Wilcoxon procedure is a test of stochastic ordering, it is particularly powerful in detecting differences between group means. As a rank-based procedure, it is not useful in testing for differences in scale (variance).

The Wilcoxon test is nonparametric. This means that it preserves the Type I error rate (i.e., false positive rate) to nominal alpha regardless of the population shape. This is a fundamental advantage over its parametric counterpart, the Student's t test, which relies on the normality distribution assumption.

When sampling from nonnormal distributions, the Wilcoxon Rank-Sum Test is often more powerful than the t test when the hypothesis being tested is a shift in location parameter. This was suggested by the large sample property known as the asymptotic relative efficiency (ARE). The ARE of the Wilcoxon (Mann-Whitney) relative to the t test under population normality is 0.955. However, under population non-normality, the ARE of the Wilcoxon Rank-Sum Test can be as high as ∞ .

Small-sample Monte Carlo studies confirmed the comparative statistical power advantage of the Wilcoxon Rank-Sum Test over the t test for departures from nonnormality. It is often three to four times more

powerful for sample sizes and treatment effect sizes common in education and psychology.

Because of the relationship between statistical power and sample size, research studies may be designed with considerably fewer participants when using the Wilcoxon Rank-Sum Test instead of the t test. It provides a considerable efficiency advantage in terms of cost, time, and effort in conducting an experiment.

In order to compute the Wilcoxon Rank-Sum Test, combine the two samples, order the scores from lowest to highest, and keep track of the score's group membership. The ordered scores are assigned ranks. If there are tied values, the average of the ranks is assigned to each of the tied scores. The Wilcoxon formula is

$$S_n = \sum_{i=1}^n R_i, \quad (1)$$

where R_i are the ranks and S_n is the sum of the ranks for a sample of size n .

The rank-sum statistic can be converted to a Mann-Whitney U in order to use commonly available tabled critical values for the U statistic. The formula for the conversion is

$$U_n = S_n - \frac{1}{2}n(n+1). \quad (2)$$

If a table of critical values is not available, the Wilcoxon Rank-Sum Test can be evaluated with a large-sample approximation using the following formula:

$$z = \frac{S_n - \frac{n(n+m+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}, \quad (3)$$

where m and n are the sample sizes of the two groups and S_n is the rank sum for sample n . The large-sample approximation for the critical value to test the obtained Mann-Whitney U statistic is

$$z = \frac{U + \frac{1}{2} - \frac{1}{2}mn}{\sqrt{\frac{mn(m+n+1)}{12}}}. \quad (4)$$

The large-sample approximation maintains the Type I error rate at nominal α when each sample size is greater than 14 for $\alpha = .05$, and greater than 29 for $\alpha = .01$, unless the data are heavily tied. In this case, each sample should be at least $n = 17$ ($\alpha = .05$) and $n = 44$ ($\alpha = .01$).

If there are few ties, they can be handled using average ranks. However, for heavily tied data, it is advisable to use a tie correction formula. The following expression is used to replace the denominator in the large sample approximation formula, either Equation 3 or Equation 4, as appropriate:

$$\sqrt{\frac{mn}{12N(N-1)}} \left[N(N^2 - 1) - \left(\sum \tau_i^3 - \sum \tau_i \right) \right] \quad (5)$$

where τ_i is the number of ties at the i th tied score.

Example

Twenty students are randomly assigned into Group 1 (treatment group) or Group 2 (comparison group). Group 1 receives a new curriculum, whereas Group 2 receives the traditional curriculum. Their test scores are displayed in Table 1. The ranked scores are presented in Table 2. For ease of computation, Group 1 is referred to as A and Group 2 is referred to as B in this table.

The sum of the ranks for Group 1 is 133, and the sum of the ranks for Group 2 is 77. For convenience, the smaller of the rank sums is taken as the obtained value and compared with the critical value. Because tabulated critical values are not often available, this value is converted to a z score based on Equation 3.

Table 1 Scores for Groups 1 and 2

Group 1	Group 2
21	18
33	19
56	20
69	32
70	35
71	54
74	55
75	68
76	72
77	73

Table 2 Sorted Data for Groups 1 and 2 With Original Ranks

Score	Group	Wilcoxon	Mann-Whitney	
		Rank	U_1	U_2
18	B	1		0
19	B	2		0
20	B	3		0
21	A	4	3	
32	B	5		1
33	A	6	4	
35	B	7		2
54	B	8		2
55	B	9		2
56	A	10	7	
68	B	11		3
69	A	12	8	
70	A	13	8	
71	A	14	8	
72	B	15		6
73	B	16		6
74	A	17	10	
75	A	18	10	
76	A	19	10	
77	A	20	10	

The result is $z = -2.12$. The p value associated with this z is 0.034. Therefore, we can conclude that the new curriculum produced statistically significant higher scores than the traditional curriculum at the $\alpha = .05$ level. In comparison, Student's t test for these data yields $t = 1.87$. With $df = 18$, $p = .078$. Thus, had the parametric t test been conducted, we would have concluded incorrectly that the new curriculum was no more effective than the traditional curriculum.

Mann-Whitney U Test

The Mann-Whitney U statistic is calculated as follows. For each score in Group 1, count the number of Bs that precede it, or count the number of As that precede each Group 2 score.

$$U_1 = 3 + 4 + 7 + 8 + 8 + 8 + 10 + 10 + 10 + 10 \\ = 77.$$

$$U_2 = 0 + 0 + 0 + 1 + 2 + 2 + 2 + 3 + 6 + 6 \\ = 22.$$

The U test is significant if the larger U_1 (77) is *greater* than the critical value or if the smaller U_2 (22) is *less* than the critical value. Tables of critical values are commonly available. They are constructed for convenience based on whichever value (U_1 or U_2) is smaller, which in this case is $U_2 = 22$. The critical value for a one-sided test with $n_1 = n_2 = 10$ and $\alpha = .05$ is 27. Because the obtained value 22 is *less* than the critical value of 27, the null hypothesis is rejected.

SPSS

To compute the Wilcoxon Rank-Sum Mann-Whitney U test via SPSS, enter the data into a column. In the next column, indicate if the score's membership is Group 1 or Group 2. Then, choose Analyze | Nonparametric Tests | 2 Independent Tests.

Highlight the scores variable and click on the arrow to move it to the Test Variable List. Highlight the grouping variable and click on the arrow to move it to the Grouping Variable. Click on Define Groups to identify Groups 1 and 2. Finally, click on OK. The output appears as Figure 1.

Ranks

	Group	N	Mean Rank	Sum of Ranks
Score	1.00	10	13.30	133.00
	2.00	10	7.70	77.00
	Total	20		

Test Statistics(b)

	Score
Mann-Whitney U	22.000
Wilcoxon W	77.000
Z	−2.117
Asymp. Sig. (2-tailed)	.034
Exact Sig. [2*(1-tailed Sig.)]	.035(a)

a. Not corrected for ties.

b. Grouping Variable: Group

Figure 1 SPSS Output for the Wilcoxon W and Mann-Whitney U Test

—Shlomo S. Sawilowsky

See also Inferential Statistics; *t* Test for Two Population Means

Further Reading

- Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5(4), 309–335.
- Fahoome, G. (2002). Twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods*, 1(2), 248–268.
- Fahoome, G., & Sawilowsky, S. (2000, April). *Twenty non-parametric statistics*. Paper presented at the annual meeting of the American Educational Research Association, SIG/Educational Statisticians, New Orleans, LA.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.
- Sawilowsky, S. (2005). Misconceptions leading to choosing the *t* test over the Wilcoxon Mann-Whitney test for shift in location parameter. *Journal of Modern Applied Statistical Methods*, 4(2), 598–600.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, 111, 353–360.
- Sawilowsky, S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.

MARKOV, ANDREI ANDREEVICH (1856–1922)

Much of what is known regarding Andrei Andreevich Markov comes from two sources. The first is a short biography provided by Ahiezer and Volkobyskii, which was written in Russian 25 years after Markov's death. The second resource is an obituary written by Vladimir A. Stekloff (or Steklov, 1864–1927) and Ya. V. Uspenskii.

Markov was born in Ryazan, Russia. He had two sons from two different marriages, both of whom were mathematicians, but especially noted was the first son, Andrei, Jr. (1903–1979). Markov completed his undergraduate work in 1878 and his master's degree in 1880 as a student of A. N. Korkin (d. 1908) and Egor Ivanovich Zolotarev (1847–1878). That same year, Markov accepted a lecturer's position in mathematics at St. Petersburg University. Four years later, he

defended his doctoral dissertation, “On Certain Applications of Algebraic Continuous Functions,” under the tutelage of Pafnuty Lvovich Chebyshev (or Tchebichef, 1821–1894).

Markov rose through the professorial ranks from Adjunct (1886), to Extraordinary (1890), and to Ordinary Academician (1896) of the Imperial Academy of Science of St. Petersburg, which was later renamed the Russian Academy of Sciences after the revolution of 1917. Two of his most well-known students were Georgy F. Voronoy (or Voronoi, 1868–1908) and Stanislaw Zaremba (1863–1942). In 1905, Markov retired from the university with the title of Distinguished Professor, but continued lecturing until his death.

Markov worked on analysis, approximation theory, converging series, continuous fractions, integrals, interpolation, number theory, limits, and probability theory. Building on the work of Chebyshev, Markov made major advances in the methods of moments in probability theory. Two of his primary pedagogical works were “Calculus of Finite Differences” and “Calculus of Probabilities.”

Markov is best known for his work on extensions from the law of large numbers that led to the development of *Markov chains*. It is a set of finite or discrete states (e.g., on vs. off) and an associated matrix that determines the probability of moving from one state to another. The primary feature, called the *Markov property*, is that the future state is determined by a random process based on the present state, but independent from all previous states, meaning that it has no memory.

A *Markov process* may be discrete or continuous, such as displacement over continuous time. A Markov process of the n th order means that both the memory and future probability state of the process are fully articulated by n elements. A *Markov field* pertains to multidimensional space.

A simple example of a Markov chain is a random walk, such as Brownian motion. The Markov chain is the progenitor and special case of stochastic processes, and led to the application of Liouville’s theorem (Joseph Liouville, 1809–1882) to the ergodic hypothesis.

—Shlomo S. Sawilowsky and Boris Shulkin

See also Markov Chain Monte Carlo Methods

Further Reading

- Higgins, J. J., & Keller-McNulty, S. (1995). *Concepts in probability and stochastic modeling*. Belmont, CA: Duxbury.
- Markov, A. A. (1948). *Izbrannye trudy po teorii nepreryvnyh drobei i teorii funktsii naimenее uklonyayushchih-sya ot nulya*. N. I. Ahiezer (Biograficheskii ocherk I primechaniya), & L. I. Volkobyskii. Moskva, Leningrad, SSSR: OGIZ, Gosudarstvennoe izdatel’svo tekhniko-teoreticheskoi literatury. [Selected works on theory of continuous fractions and theory of the least deviated from zero functions (N. I. Ahiezer, Ed., & L. I. Volkobyskii, Trans., selected chapters). Moscow, Leningrad, USSR: OGIZ, State Publishing House for Technical-Theoretical Literature.]
- Sawilowsky, S. S. (2004). A conversation with R. Clifford Blair on the occasion of his retirement. *Journal of Modern Applied Statistical Methods*, 3(2), 518–566.

MARKOV CHAIN MONTE CARLO METHODS

Markov chain Monte Carlo (MCMC) is a modification of the Monte Carlo simulation method. In a typical Monte Carlo simulation, we are trying to compute the expectation of some random variable, which we do by sampling and using the law of large numbers. However, often, it is either impossible or intractable to obtain enough independent samples from the desired distribution. The basic modification in MCMC from Monte Carlo is that in MCMC we use dependent samples, which are generated by a Markov chain. Under suitable and very general conditions, averages computed from the trajectory of a Markov chain will also converge to the desired expectation.

Because MCMC was motivated by Monte Carlo simulation methods, we start with a short discussion of Monte Carlo methods.

Monte Carlo Simulation

The basic problem in Monte Carlo simulations is to estimate a target density π on some (usually very complicated and large) state space. This estimation is done by drawing independent and identically distributed samples from π and using the empirical distribution as an approximation to the true distribution.

Clearly, the central concern in a Monte Carlo simulation is drawing a random sample from the target distribution π . Because of this, many sampling methods have been devised.

Rejection Sampling

Many MCMC algorithms are strongly related to rejection sampling, and this is the main reason for our discussion of rejection sampling.

A simple example best illustrates the idea behind rejection sampling. Let π is a distribution on $[0,1]$ with density $p(x)$. Now, suppose that there is another distribution with density $q(x)$ on $[0,1]$ such that $p(x) < Mq(x)$ for all x and such that we can sample from q . To get one sample X from π , let Y be a sample from q and U be a sample from $U[0,1]$. If $U < p(Y)/(Mq(Y)) = p(Y)/M$, then we accept Y as a sample from q and set $X = Y$, else we reject and try again (see Figure 1 for an illustration of this).

We see that the infinitesimal probability that $X = x$ is returned is

$$\begin{aligned} Pr(X = x)dx &= \frac{Pr(Y = x \text{ and } x \text{ is returned})}{Pr(\text{something is returned})}dx \\ &= \frac{q(x)dx \frac{p(x)}{Mq(x)}}{\int_x q(x) \frac{p(x)}{Mq(x)} dx} = p(x)dx, \end{aligned}$$

so this gives a perfect sample from $p(x)$. Notice, however, that the efficiency is $1/M$, because only one out of every M samples will be accepted.

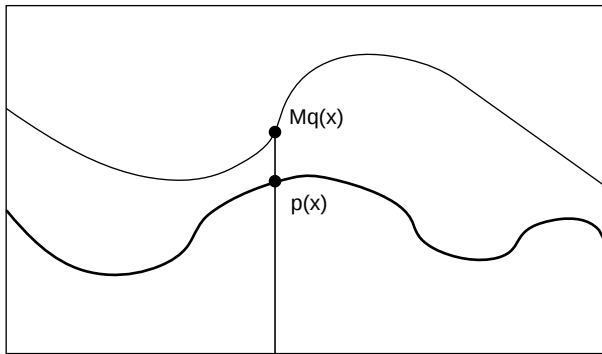


Figure 1 Rejection Sampling

In general, any distribution q with $p(x) \leq Mq(x)$ will do (not just uniform), and the efficiency will again be $1/M$.

Generalities on MCMC

For Monte Carlo simulation, we require samples from the distribution π . This is very often either impossible or highly impractical to do. Even if it is possible to generate one perfect sample, we need many independent samples for a Monte Carlo simulation. MCMC eases this problem by giving up the requirement for independent samples from π and generating instead many correlated approximate samples from π , with the approximation improving as the simulation continues.

The ergodic theorem gives some very general conditions under which MCMC will work. A Markov chain with state space Ω is *irreducible* if it is possible for the chain to get from any state in Ω to any other state in Ω , and it is *positively recurrent* if the expected time for transition from state i to state j is finite for every $i, j \in \Omega$. Clearly if Ω is finite and the chain is irreducible, then it is also positively recurrent.

Ergodic Theorem for Markov Chains

Let $\{X_n\}$ be an irreducible positively recurrent Markov chain on the state space Ω with stationary distribution π . If $f: \Omega \rightarrow \mathbb{R}$ is π -summable (in the sense that $\sum_{\omega} |f(\omega)| \pi(\omega) < \infty$), then

$$\frac{1}{N} \sum_{i=1}^N f(X_i) \rightarrow \sum_{\omega \in \Omega} f(\omega) \pi(\omega) = E_{\pi}(f).$$

Notice that the ergodic theorem also yields $P(X_n = \omega) \rightarrow \pi(\omega)$ by considering f to be the characteristic function of $\{\omega\}$.

Of course, in practice, the rate of convergence is a critical issue.

Detail Balance

Let π be a distribution on Ω . A very important way of finding a Markov chain on Ω with transition matrix

P , which has invariant distribution π , is to ensure that the *detail balance* condition

$$\pi(x)P(y | x) = \pi(y)P(x | y), \quad \text{for all } x, y \in \Omega,$$

holds. It is easy to see that this condition is sufficient; simply sum both sides over all possible $x \in \Omega$ to get

$$\sum_x P(y | x)\pi(x) = \pi(y).$$

Markov chains that satisfy the detail balance condition are called *reversible* chains.

Metropolis Sampler

In 1953, Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller published the paper “Equations of State Calculations by Fast Computing Machines,” thereby introducing to the world the incredibly useful Metropolis simulation algorithm. This algorithm has been cited as among the top 10 algorithms having the greatest influence on the development of science and engineering.

The basic idea, like many great ideas, is rather simple. Given a desired distribution π on Ω , we split the chain into a *proposal* phase and an *acceptance* phase (like rejection sampling). For the proposal phase, the Metropolis algorithm uses a symmetric irreducible Markov chain Q (so $Q(x | y) = Q(y | x)$). Given the current state x , we use Q to generate a new state y and accept y with probability

$$A(y | x) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\}.$$

Thus, the transition matrix P for the Metropolis algorithm is a product of the form $P(y | x) = Q(y | x)A(y | x)$ for $y \neq x$ (the diagonal terms of P are forced by the fact that P is a transition matrix). From this, we see that (without loss of generality, $\pi(y) \leq \pi(x)$),

$$\begin{aligned} \pi(x)P(y | x) &= \pi(x)Q(y | x)\frac{\pi(y)}{\pi(x)} = \pi(y)Q(y | x) \\ &= \pi(y)Q(x | y) = \pi(y)P(x | y), \end{aligned}$$

so the Metropolis chain satisfies detail balance and has π as invariant distribution.

The Metropolis algorithm will produce many repeated states in the chain X_n . In computing expectations using this chain, it is important to keep these repeats because this is one way the algorithm corrects the bias from the proposal process, thus allowing it to converge to π .

The separation of the proposal and acceptance phases makes designing a chain much simpler, because the proposal process can be virtually any process. The choice of proposal process is often guided by the structure of the state space. The condition that Q be irreducible and symmetric is rarely much of a problem in practice. For Ω finite, the proposal process is often defined by first giving a local (symmetric) neighborhood structure and then choosing y uniformly from the neighbors of the current x , and the resulting Q is irreducible as long as the given neighborhood system is reasonable. We are free to let the structure of the problem guide the neighborhood system (or proposal process) and use the acceptance phase to correct the proposal process in such a way as to ensure the desired stationary distribution.

Another very important feature of the Metropolis algorithm is that one need know π only up to a multiplicative factor, because only ratios are used in computing the acceptance probability. This can be very important in situations where we want specific relative probabilities but the state space is too large or complicated to allow us to compute a normalizing constant. Optimization by simulated annealing is often such a situation.

Simulated Annealing for Global Optimization

Given an energy function E on the state space Ω and a temperature $T > 0$, the *Boltzmann distribution* is the probability distribution

$$\pi(x) = \frac{e^{-E(x)/T}}{Z_T},$$

where Z_T is a normalizing constant called the partition function. Clearly, this distribution peaks at the states with the lowest temperature, and this peaking becomes more pronounced as $T \rightarrow 0$. Given this fact, one way to find the minima of the energy E might be to try to lower the temperature in a controlled way (clearly, setting $T = 0$ won't work). This is the basic idea of simulated annealing.

Simulated annealing typically uses the Metropolis sampler, with corresponding acceptance

$$A(y | x) = \min\{1, e^{(E(y) - E(x))/T}\},$$

where we notice that the partition function has canceled out. This is a clear advantage to using the Metropolis algorithm, because the partition function is almost always impossible to compute.

The interpretation of the Metropolis acceptance in this context is that if the energy of the proposed state is smaller than the energy of the current state, we always accept the transition, whereas if the energy of the proposed state is greater than the energy of the current state, we accept the transition only with some probability that decreases to zero as T decreases to zero. Care must be taken in lowering the temperature that it is not lowered too quickly, or the process will “freeze.” Conceptually, one fixes the temperature at some level, allows the Markov chain to approach the Boltzmann distribution for that temperature, and then lowers the temperature. At lower temperatures, it takes longer for the Markov chain to converge to the Boltzmann distribution (for that temperature). In practice, one usually lowers the temperature on each iteration of the chain (so that the chain is a nonhomogeneous Markov chain with transition matrix depending on time), but following a very slow cooling schedule. Theoretical results indicate that a cooling schedule of the form

$$T = \frac{c}{\log(\text{iteration count})}$$

is sufficiently slow to guarantee convergence to the minimum energy states. Because this cooling schedule is incredibly slow, most actual simulated annealing algorithms use a faster cooling schedule.

Metropolis-Hastings

W. K. Hastings's 1970 paper generalized the work of Metropolis et al. and introduced the Metropolis-Hastings algorithm. This algorithm is the most frequently used algorithm in MCMC, and most of the other algorithms are instances or modifications of it.

The Metropolis-Hastings algorithm also splits the chain into a proposal and acceptance phase. However, the proposal process Q is not assumed to be symmetric, but it does have to satisfy $Q(x | y) > 0$ whenever $Q(y | x) > 0$ and be irreducible. This time, the acceptance probability is

$$A(y | x) = \frac{S(y | x)}{1 + \frac{\pi(x)Q(y|x)}{\pi(y)Q(x|y)}} = \frac{S(y | x)}{1 + T(y | x)},$$

where $S(y | x) = S(x | y)$ is any symmetric Markov kernel. Using these definitions it is easy to see that $P(y | x) = Q(y | x)A(y | x)$ satisfies detail balance for π . The advantages of the Metropolis-Hastings algorithm over the Metropolis algorithm are that Q need not be symmetric and the flexibility in choosing S .

Because $A(y | x) \in [0, 1]$, we must have that $S(y | x) \leq 1 + \min\{T(y | x), T(x | y)\}$. The Metropolis algorithm corresponds to the case of equality in this inequality. In fact, over the class of all possible Metropolis-Hastings transition matrices, the Metropolis algorithm has the maximum chance of accepting a move (so it minimizes the diagonal elements). A theorem of P. Peskun then implies that the asymptotic variance of the Metropolis algorithm is optimal (in the sense of being smallest) among the class of all Metropolis-Hastings chains with a fixed proposal process. Thus, even though the general Metropolis-Hastings algorithm gives more flexibility in the design of the chain, the Metropolis algorithm is, in some ways, optimal when it is feasible.

Another special case of the Metropolis-Hastings sampler is Barker's algorithm, where we choose $S(y | x) = 1$ and get the nice acceptance probability

$$A(y | x) = \frac{\pi(y)Q(y | x)}{\pi(x)Q(x | y) + \pi(y)Q(y | x)}.$$

Another useful modification to the generic Metropolis-Hastings sampler is the Langevin Metropolis-Hastings algorithm. The generic Metropolis-Hastings sampler uses the acceptance phase to correct the bias from the proposal process. This often results in many rejections and a slow convergence to π . Partly, this is due to the fact that the proposal process can be independent of π . The Langevin Metropolis-Hastings is motivated by a stochastic differential equation and introduces a bias into the proposal process that is influenced by the derivative of the density of π , which often speeds up the convergence.

Gibbs Sampler

The Gibbs sampler is especially designed for situations where X_n is multivariate and we can easily (or more easily) sample from the marginal distributions than from the full distribution. This is best explained by the following example.

Take a 10×10 grid with each site being in the state 0 or in the state 1 (see Figure 2 where we shade all 1 states). For each site, we consider the four adjacent sites (with wraparound for the edges) and count the number of neighbors in state 0. We want to have the marginal distributions given in Table 1. The problem is to sample a “random” 10×10 grid that has these local marginal distributions.

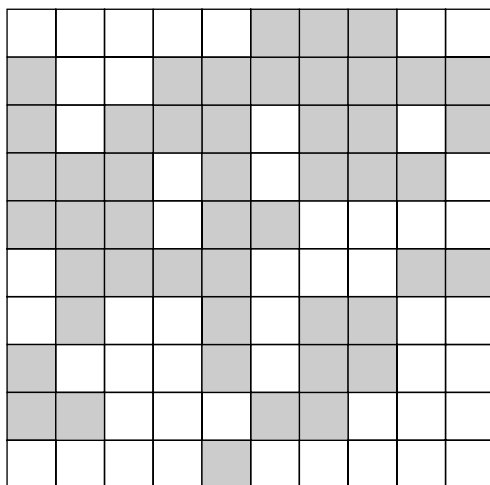


Figure 2 Simple Grid for Gibbs Sampler Example

Table 1 Marginal Distributions for Gibbs Sampler Example

# 0 neighbors	Prob. of being 0
0	.1
1	.3
2	.5
3	.7
4	.9

The Gibbs sampler solves this problem by the following steps in cycles. First, we randomly initialize the states in the grid, and then for each cycle do the following:

- Choose a (random) ordering of the grid points.
- For each grid point in turn in the order chosen, sample the value of the current grid point according to the distribution given by its four neighbors.

Sometimes, we keep a copy of the “old” grid around for the marginal draws and update the entire grid only at the end of the sample. It is also possible to simplify the algorithm by eliminating the cycles and randomly choosing a site to update at each iteration. These different choices shouldn’t affect the limiting distribution but might influence the speed of convergence. The grid in Figure 2 was generated using this model.

The Gibbs sampler can also incorporate a Metropolis or Metropolis-Hastings proposal/acceptance scheme.

Other Methods and Discussion

There are many variations on these basic samplers and other general Monte Carlo and MCMC algorithms. Among these, we mention the EM algorithm (an algorithm for ML and MAP point estimates), the Slice sampler (a generalized Gibbs sampler), Reversible Jump MCMC, and Sequential Monte Carlo.

A useful property of MCMC is the ability to combine several samplers. If P_1 and P_2 are transition kernels that both have π as an invariant distribution, then so do $\lambda P_1 + (1 - \lambda)P_2$ (for $0 \leq \lambda \leq 1$) and $P_1 P_2$. The

Gibbs sampler can be thought of as an example of the second kind of combination, where each factor changes only one component of X at a time.

Obtaining good estimates of the convergence rate is usually difficult but extremely useful. One major problem in running an MCMC simulation is the problem of how long is long enough. There is a large (and growing) literature on bounding the mixing rate of a Markov chain, and these estimates can give useful guidelines on the required length of a run. In simple cases, the running time is shown to be polynomial in the dimension of the state space.

A recent and very exciting development in MCMC is the existence of algorithms for perfect sampling. These completely answer the question of how long you need to run the MCMC chain. Two such algorithms are Coupling From the Past and Fill's Algorithm. Although they are limited to certain types of Markov chains, the class of possible chains is steadily growing, and in these instances, these algorithms guarantee a perfect draw from π . Essentially, these two algorithms provide a computable stopping criterion.

—Franklin Mendivil

See also EM Algorithm; Markov, Andrei Andreevich; Monte Carlo Methods; Simulated Annealing

Further Reading

- Besag, J. (2001). *Markov chain Monte Carlo for statistical inference*. Center for Statistics and the Social Sciences Working Paper No. 9. Available at <http://www.csss.washington.edu/Papers/>
- Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. New York: Springer.
- Fill, J. (1988). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 8, 131–162.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Givens, G., & Hoeting, J. (2005). *Computational statistics*. New York: Wiley Interscience.
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

- Kirkpatrick, S., Gelatt, G., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Peskun, P. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60, 607–612.
- Propp, J., & Wilson, B. (1998). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9, 223–252.
- Wilson, D. (2000). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). *Journal of the American Statistical Association*, 92, 49–60.

MCMC preprints page: <http://www.statslab.cam.ac.uk/~mcmc/>
 Perfectly random sampling with Markov chains: <http://http://dbwilson.com/exact/>

MATRIX ANALOGIES TEST

The Matrix Analogies Test (published by Harcourt Assessment) consists of two forms: the Matrix Analogies Test–Expanded Form (MAT-EF) and the Matrix Analogies Test–Short Form (MAT-SF). Stimulus items on both tests employ abstract designs of the standard progressive matrix type. The matrix design of the stimulus items of both the MAT-EF and the MAT-SF require minimal verbal comprehension and no verbal response on the part of the examinee.

The MAT-EF is an individually administered test of nonverbal reasoning ability designed for children and adolescents 5 to 17 years old. The MAT-EF is reported to be appropriate for assessing the abilities of children with learning disabilities, mental retardation, hearing or language impairments, physical disabilities, those who may speak more than one language, and those with the ability to perform at the gifted level. It may be used as a stand-alone measure or as part of a comprehensive battery of tests.

The MAT-EF is organized into four specific item groups (Pattern Completion, Reasoning by Analogy, Serial Reasoning, and Spatial Visualization). The author claimed that these item groups were developed based on the results of item factor analysis. The MAT-EF was designed in such a way as to reduce the effects of impaired color vision. The colors (blue, yellow, black, and white) were used in the stimulus materials to reduce the influence of impaired color vision.

When administering the MAT-EF, a maximum of 12 minutes is allowed for each item group. The MAT-EF yields an overall standard score ($M = 100$, $SD = 15$) and item-group standard scores ($M = 10$, $SD = 3$). The MAT-EF is reported to yield similar scores for boys and girls, whites and African Americans, and Hispanics and non-Hispanics. Additionally, research has shown that similar scores are obtained across samples from different countries.

The MAT-SF is a group-administered version of the MAT-EF that is also used to obtain a measure of nonverbal reasoning ability. The MAT-SF consists of 34 items. It yields stanines and percentile scores for the total test score. As with the MAT-EF, the MAT-SF requires minimal motor involvement and a minimal level of verbal comprehension. The MAT-SF is appropriate for use with the same age group as the MAT-EF. The primary use of the MAT-SF is as a quick screening device to help identify students who may be at risk and students who may be gifted and are performing above age expectations. The MAT-SF uses self-scoring answer sheets that eliminate the need for scoring keys and provide immediate results.

—Thomas O. Williams, Jr.

$$A = \begin{bmatrix} 0 & \frac{-3}{2} & 5 \\ 1 & -7 & 2 \end{bmatrix}, M = \begin{bmatrix} 3 & 2 \\ -\pi & 0 \\ 1 & 1 \end{bmatrix}, S = \begin{bmatrix} 2 & -1 & 3 & 0 \\ 0 & 2 & -1 & 3 \\ 3 & 0 & 2 & -1 \\ -1 & 3 & 0 & 2 \end{bmatrix}$$

Given an m by n matrix A , for each $i = 1, \dots, m$, and $j = 1, \dots, n$, the symbol a_{ij} denotes the entry of A in row i , column j .

For m by n matrices A and B , their sum, $A + B$, is the m by n matrix such that for all $i = 1, \dots, m$, $j = 1, \dots, n$, the entry in row i , column j is given by $a_{ij} + b_{ij}$. The sum of two matrices is defined only when both the number of rows and the number of columns of the summands agree. For example,

$$\begin{bmatrix} 0 & \frac{-3}{2} & 5 \\ 1 & -7 & 2 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & 1 & -3 \\ -2 & 6 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{-1}{2} & 2 \\ -1 & -1 & 2 \end{bmatrix}.$$

For any m by n matrix A and any scalar x , the scalar multiple xA is the m by n matrix such that for all $i = 1, \dots, m$, $j = 1, \dots, n$, the entry in row i , column j is given by xa_{ij} . For example, if

$$A = \begin{bmatrix} 0 & \frac{-3}{2} & 5 \\ 1 & -7 & 2 \end{bmatrix}, \text{ then } -4A = \begin{bmatrix} 0 & 6 & -20 \\ -4 & 28 & -8 \end{bmatrix}.$$

See also Culture Fair Intelligence Test; Intelligence Tests; Matrix Operations; Raven's Progressive Matrices

Further Reading

- Naglieri, J. A. (1985). *Matrix Analogies Test—Expanded Form*. San Antonio, TX: The Psychological Corporation.
- Naglieri, J. A. (1985). *Matrix Analogies Test—Short Form*. San Antonio, TX: The Psychological Corporation.
- Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. San Antonio, TX: Psychological Corporation.
- Naglieri, J. A. (2003). Naglieri Nonverbal Ability Tests: NMAT and MAT-EF. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 175–189). New York: Kluwer Academic/Plenum.

Harcourt Assessment: <http://www.harcourtassessment.com>

J. A. Naglieri Web site: <http://www.mypsychologist.com>

MATRIX OPERATIONS

A matrix is a rectangular array of numbers, and is called m by n when it has m rows and n columns. Examples include the matrices A (which is 2 by 3), M (3 by 2), and S (4 by 4) below:

Addition and scalar multiplication can be combined for two m by n matrices A and B , so that the matrix $xA + yB$ has the entries $xa_{ij} + yb_{ij}$ for all $i = 1, \dots, m, j = 1, \dots, n$. Thus, if

$$A = \begin{bmatrix} 0 & \frac{-3}{2} & 5 \\ 1 & -7 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} \frac{1}{2} & 1 & -3 \\ -2 & 6 & 0 \end{bmatrix}, \text{ then } 3A - 2B = \begin{bmatrix} -1 & \frac{-13}{2} & 21 \\ 7 & -33 & 6 \end{bmatrix}.$$

For an m by n matrix U and an n by k matrix V , the matrix product UV is the m by k matrix such that for each $i = 1, \dots, m, j = 1, \dots, k$, the entry in row i , column j of UV is given by $\sum_{p=1}^n u_{ip}v_{pj}$. The matrix product UV is defined only when the number of columns of U agrees with the number of rows of V . For example, if

$$U = \begin{bmatrix} 1 & -1 & 2 \\ 3 & 0 & 1 \end{bmatrix}, V = \begin{bmatrix} 2 & -1 \\ -3 & 4 \\ 0 & -2 \end{bmatrix}, \text{ then } UV = \begin{bmatrix} 5 & -9 \\ 6 & -5 \end{bmatrix} \text{ whereas } VU = \begin{bmatrix} -1 & -2 & 3 \\ 9 & 3 & -2 \\ -6 & 0 & -2 \end{bmatrix}.$$

If a matrix S has an equal number of rows and columns, we say that S is square. Given a square matrix S , for each natural number k , the k th power of S , denoted S^k , is the k -fold product of S with itself. For example, if

$$S = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 1 & 1 \\ -1 & 0 & 2 \end{bmatrix}, \text{ then } S^2 = \begin{bmatrix} -1 & 1 & 1 \\ -2 & 0 & 3 \\ -2 & -1 & 4 \end{bmatrix}, \text{ and } S^3 = \begin{bmatrix} -2 & 0 & 3 \\ -3 & -2 & 6 \\ -3 & -3 & 7 \end{bmatrix}.$$

For each natural number n , the identity matrix of order n , denoted by I_n , is the n by n matrix with diagonal entries equal to 1 and all other entries equal to 0. For example,

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For any m by n matrix A and any n by k matrix B , we have $AI_n = A$ and $I_n B = B$. Given an n by n matrix S , we say that S is invertible if there is an n by n matrix T such that $ST = I_n$ and $TS = I_n$. If such a T exists, it is the inverse of S and is denoted by S^{-1} . For any natural number k , the k -fold product of S^{-1} with itself is denoted S^{-k} . For each natural number k , we have $(S^k)^{-1} = S^{-k}$. For example if

$$S = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 1 & 1 \\ -1 & 0 & 2 \end{bmatrix}, \text{ then } S^{-1} = \begin{bmatrix} 2 & -2 & 1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix} \text{ and } S^{-2} = \begin{bmatrix} 3 & -5 & 3 \\ 2 & -2 & 1 \\ 2 & -3 & 2 \end{bmatrix}.$$

Because of their computational utility and well-developed theory, matrices arise throughout the mathematical sciences and have numerous applications in science and engineering.

—Steve Kirkland

See also Eigenvalues

MAXIMUM LIKELIHOOD METHOD

The maximum likelihood method is applicable to any scientific problem in which it is desired that unknown or unobservable quantities, called *parameters*, be estimated based on observed data. Thus, maximum

likelihood is a method of solving the statistical problem of *estimation*.

Under the maximum likelihood principle, a parameter is estimated by the value that maximizes the likelihood of the observed data. In other words, the maximum likelihood estimate (MLE) of a parameter is the value that gives the observed data the highest probability possible. Depending on the complexity of the underlying model, MLEs may be solvable by elementary calculus, or they may require sophisticated computational methods. The former instance is illustrated in the following example.

In a survey of 6,000 randomly selected U.S. high school students, 2,166 knew that Geoffrey Chaucer wrote *The Canterbury Tales*. What is the MLE for the percentages of *all* U.S. high school students who know this fact?

Let n denote the number of subjects surveyed, and X the number who knew the correct answer. Let p denote the proportion of all U.S. high school students who know the answer. The sample size n and population percentage p are fixed quantities, or parameters, the former known and the latter unknown. The observed quantity X is a random variable—it can assume any integer value between 0 and n , depending on the result of the random sample—and follows the binomial distribution. For any realizable value x , the probability is given by

$$\Pr(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad (1)$$

where $k! = k(k-1) \cdots 2 \cdot 1$ for any integer k .

The MLE of p , denoted by $\hat{p}$, is the number between 0 and 1 that maximizes the value of Equation 1. In Equation 1, the quantity p is held fixed, and a probability is evaluated for any possible value of x from 0 to n . The method of maximum likelihood involves fixing x at the observed value and considering Equation 1 a function of the unknown success probability p . Consider the Chaucer question. Based on the results of the survey, the values $p = 0$ and $p = 1$ can be eliminated, but any number between 0 and 1 remains a possibility. Here, $n = 6,000$ and $x = 2,166$; the probability of the observed result is given as follows for selected values of p .

p	$\Pr(X = 2166) \text{ where } X \sim \text{Binomial}(6000, p)$
.35	.00219
.36	.01058
.37	.00377

This probability is maximized at $p = .361$. More generally, elementary calculus will show that the MLE is given by $\hat{p} = x/n$; hence, the MLE of the population percentage is just the sample percentage. Thus, by the maximum likelihood method, it is estimated that 36.1% of students know that Chaucer wrote *The Canterbury Tales*.

Of course, few problems admit such straightforward solutions. Indeed, the development of efficient algorithms for finding MLEs is an active area of research.

—Ronald Neath

See also Binomial Distribution/Binomial and Sign Tests; Parameter

Further Reading

- Booth, J., Hobert, J., & Jank, W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1, 333–349.
- Rice, J. A. (1994) *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Method of maximum likelihood: http://en.wikipedia.org/wiki/Maximum_likelihood

MCNEMAR TEST FOR SIGNIFICANCE OF CHANGES

It is often of interest to examine changes in the dichotomous categorical responses taken from subjects before and then after some treatment condition is imposed (i.e., evaluating repeated measurements of the same subjects using them as their own controls). In 1947, the psychologist Quinn McNemar developed a simple and valuable technique for comparing differences between the proportions in the responses before and after. McNemar's procedure is the categorical

data counterpart of the t test for the mean difference in matched, paired, or related samples.

McNemar's procedure has enjoyed widespread usage in both behavioral and medical research and some attention in business, particularly with applications in advertising or marketing research, wherein it may be desirable to evaluate the significance of changes in attitudes and opinions.

Development

The dichotomous responses from a sample of n' individuals over two periods of time may be tallied into a 2×2 table of cross-classifications as follows:

With respect to the population from which the aforementioned sample was taken, let p_{ij} be the probability of responses to the i th category before the treatment condition was imposed and the j th category after. The pairs of marginal probabilities before and after treatment sum to unity; that is, $p_{1\cdot} + p_{2\cdot} = 1$ and $p_{\cdot 1} + p_{\cdot 2} = 1$.

Table 1 2×2 Table of Cross-Classifications for a Sample of n' Subjects

$I \backslash II$	+	−	Totals
+	x_{11}	x_{12}	$x_{1\cdot}$
−	x_{21}	x_{22}	$x_{2\cdot}$
Totals	$x_{\cdot 1}$	$x_{\cdot 2}$	n'

where

I = Time period I (before treatment) in a repeated measurements experiment

II = Time period II (after treatment) in a repeated measurements experiment

+ = positive classification

− = negative classification

x_{11} = number of subjects with a positive response both before and after treatment

x_{12} = number of subjects with a positive response before treatment and a negative response after treatment

x_{21} = number of subjects with a negative response before treatment and a positive response after treatment

x_{22} = number of subjects with a negative response both before and after treatment

n' = sample size

Testing for Significance of Changes in Related Proportions

In order to investigate changes in repeated dichotomous measurements, the null hypothesis is that of symmetry:

$$H_0: p_{ij} = p_{ji} \text{ for } i \neq j.$$

That is, the null hypothesis tested is conditioned on those $n = x_{12} + x_{21}$ individuals whose responses change, where the probability (p_{21}) of a switch to a more favorable position is equal to the probability (p_{12}) of a switch to a less favorable position, and that this probability is 0.5.

Under the null hypothesis the random variable x_{12} is binomially distributed with parameters n and 0.5, as is the random variable x_{21} . The expected value of each of these binomial distributions is $0.5n$, and the variance for each is $0.25n$. McNemar's procedure enables an exact test of the null hypothesis using the binomial probability distribution with parameters n and $p = 0.5$.

The McNemar test statistic M , written as $\text{Min}[x_{12}, x_{21}]$, is defined as the minimum of the response tallies x_{12} or x_{21} in the cross-classification table.

For a two-tailed test, the null hypothesis can be rejected at the α level of significance if

$$P(M \leq \text{Min}[x_{12}, x_{21}] | n \text{ and } p_{12} = p_{21} = 0.5) = \sum_{M=0}^{\text{Min}[x_{12}, x_{21}]} \frac{n!}{M!(n-M)!} (0.5)^n \leq \alpha/2.$$

For a one-tailed test, the null hypothesis may be rejected if

$$P(M \leq \text{Min}[x_{12}, x_{21}] | n \text{ and } p_{12} = p_{21} = 0.5) = \sum_{M=0}^{\text{Min}[x_{12}, x_{21}]} \frac{n!}{M!(n-M)!} (0.5)^n \leq \alpha.$$

Using Microsoft Excel's BINOMDIST function, exact p values of the McNemar test are obtained for all $n \leq 1,000$.

For studies where $n > 1,000$ a simple normal approximation formula for the test statistic M is given by

$$Z = \frac{M - 0.5n}{0.5\sqrt{n}},$$

where $Z = N(0,1)$, the standardized normal distribution. The decision rule for testing the null hypothesis ($H_0: p_{ij} = p_{ji}$ for $i \neq j$) depends on whether the test is two-tailed or one-tailed. Based on its definition, M cannot exceed $0.5n$, so the test statistic Z can be rejected only in the left tail of a standardized normal distribution. For a two-tailed test against the alternative $H_1: p_{ij} \neq p_{ji}$, the decision rule is to reject H_0 if $Z \leq Z_{\alpha/2}$. For a one-tailed test, the decision rule is to reject H_0 if $Z \leq Z_{\alpha}$.

Applying the McNemar Procedure

The following hypothetical example is presented: Suppose a consumer panel of $n' = 500$ participants is selected, and the panel members are initially asked to state their preferences for two competing Internet service providers, say, AOL versus EarthLink. Suppose 290 panelists initially say they prefer AOL. After exposing the entire panel to a newly designed advertisement as part of some intensive marketing campaign strategy for EarthLink, the same 500 panelists are again asked to state their preferences. Suppose that of the 290 panelists who previously preferred AOL, 260 maintain their brand loyalty but 30 switch to EarthLink. Moreover, suppose that of the 210 panelists who initially preferred EarthLink, 200 remain

brand loyal but 10 switch to AOL. The results are displayed in Table 2.

To test the null hypothesis of symmetry

$$H_0: p_{ij} = p_{ji} \text{ for } i \neq j$$

(that is, the marketing campaign strategy has no effect—Internet users are just as likely to shift their preference from AOL to EarthLink as they are to switch from EarthLink to AOL), it may be tested against the two-tailed alternative

$$H_1: p_{ij} \neq p_{ji}$$

(that is, exposure to the advertisement does influence one's disposition to switch from one Internet service provider to the other).

For these data, the McNemar procedure enables an exact test of the null hypothesis using the binomial probability distribution with parameters $n = x_{12} + x_{21} = 40$ and $p = 0.5$ and with a stated level of significance $\alpha/2$ in each tail. For a two-tailed test, the McNemar test statistic M —defined here as 10, or the minimum of the response tallies x_{12} and x_{21} from the cross-classification table—can be rejected at the $\alpha = 0.05$ level of significance if

$$P(M \leq 10 | n = 40 \text{ and } p_{12} = p_{21} = 0.5) \leq 0.025.$$

Using Microsoft Excel's BINOMDIST,

$$\begin{aligned} P(M \leq 10 | n = 40 \text{ and } p_{12} = p_{21} = 0.5) \\ &= \sum_{M=0}^{10} \frac{40!}{M!(40-M)!} (0.5)^{40} = 0.00111 < \alpha/2 \\ &= 0.025. \end{aligned}$$

Thus, the null hypothesis is rejected. The p value is 0.00222. The advertisement significantly increased EarthLink's market share at the expense of its arch rival AOL.

Had Microsoft Excel not been readily available, the normal approximation formula could have been employed. For these data,

Table 2 Hypothetical Results of a Marketing Campaign

<i>Before \ After</i>	<i>EarthLink</i>	<i>AOL</i>	<i>Totals</i>
<i>EarthLink</i>	$x_{11} = 200$	$x_{12} = 10$	$x_{1.} = 210$
<i>AOL</i>	$x_{21} = 30$	$x_{22} = 260$	$x_{2.} = 290$
<i>Totals</i>	$x_{.1} = 230$	$x_{.2} = 270$	$n' = 500$

$$Z = \frac{M - 0.5n}{0.5\sqrt{n}} = \frac{10 - 20}{0.5\sqrt{40}} = -3.16.$$

For a two-tailed test at the α level of significance, H_0 is rejected if $Z \leq Z_{\frac{\alpha}{2}}$. Using the traditional .05 level of significance, because $Z = -3.16 < Z_{0.025} = -1.96$, H_0 is rejected, and it is concluded that the advertisement significantly influenced Internet provider preference. A significantly greater number of users switched to EarthLink from AOL than switched from EarthLink to AOL. The p value, or size of this test, is approximated to be 0.00158, a highly significant result, but more liberal than the exact p value obtained from the binomial distribution, 0.00222.

Discussion

Testing for Matched or Paired Differences in Proportions

The McNemar test may also be used to compare differences in the dichotomous categorical responses from a set of matched or paired subjects when one member of each pair is exposed to a particular treatment condition and the other member to a different treatment condition. The null hypothesis of equality of marginal probability distributions

$$H_0: p_i = p_{.i} \text{ for } i = 1, 2$$

is equivalent to testing the hypothesis of symmetry in the 2×2 table of cross-classifications. If $p_{11} = p_{1.}$, then $p_{11} + p_{12} = p_{11} + p_{21}$ so that $p_{12} = p_{21}$ (i.e., symmetry).

Using Directional or Nondirectional Tests

Aside from justification through ethical arguments, biostatistician Joseph Fleiss proposed that a two-tailed test be used in lieu of a one-tailed test "in the vast majority of research undertakings . . . to guard against the unexpected." In the hypothetical application presented here, if the advertisement copy is seen by the user as effective, EarthLink would be expected to gain market share at the expense of AOL. However, if the advertisement is deemed unrealistic, the strategy may backfire and EarthLink could lose market share to AOL.

Forming McNemar's Confidence Interval Estimate

A $(1 - \alpha)\%$ confidence interval estimate of the differences in related population proportions ($p_{.1} - p_{1.}$) was given by Marascuilo and McSweeney as

$$(\hat{p}_{.1} - \hat{p}_{1.}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_{.1}\hat{p}_{.2}}{n'} + \frac{\hat{p}_{1.}\hat{p}_{2.}}{n'} - \frac{2(\hat{p}_{11} - \hat{p}_{1.}\hat{p}_{1.})}{n'}}$$

where $\hat{p}_{.1}$, $\hat{p}_{1.}$, $\hat{p}_{.2}$, $\hat{p}_{2.}$ are the estimators of the respective parameters, $p_{.1}$, $p_{1.}$, $p_{.2}$, $p_{2.}$.

To form a 95% confidence interval estimate of the differences in related population proportions ($p_{.1} - p_{1.}$), representing increase in support for EarthLink over AOL following an intense marketing campaign, the estimates from the hypothetical data in Table 2 are summarized as follows:

Before Marketing Campaign	After Marketing Campaign
EarthLink: $\hat{p}_{1.} = \frac{210}{500} = 0.42$	$\hat{p}_{.1} = \frac{230}{500} = 0.46$
AOL: $\hat{p}_{2.} = \frac{290}{500} = 0.58$	$\hat{p}_{.2} = \frac{270}{500} = 0.54$

For these data

$$(0.46 - 0.42)$$

$$\pm 1.96 \sqrt{\frac{(0.46)(0.54)}{500} + \frac{(0.42)(0.58)}{500} - \frac{2[(0.40) - (0.46)(0.42)]}{500}}$$

and

$$0.016 \leq (p_{.1} - p_{1.}) \leq .065$$

It can be concluded with 95% confidence that the gain in support for EarthLink at the expense of AOL as a result of the marketing campaign is between 1.6% and 6.5%.

Comments

It is essential to a good data analysis that the appropriate statistical procedure be applied to a specific situation. When comparing differences in two proportions based on related samples, the McNemar test

should always be used. Failure to do so will often lead to erroneous conclusions. A researcher unaware of the magnitude of the correlated proportions that are accounted for in the standard error term shown in the McNemar confidence interval formula may erroneously treat the paired responses as independent and thus inflate the standard error, causing a loss of precision in the confidence interval estimate or a loss in statistical power when testing inappropriately for differences in the two proportions.

The pedagogical advantage to the confidence interval approach to McNemar's procedure over the significance testing approach derives from the fact that the former makes use of all 500 repeated responses, whereas the corresponding hypothesis test statistic is conditioned on the reduced set that contains only the brand-switching panelists off the main diagonal of the cross-classifications table; the fact that the test statistic discards the brand-loyal panelists unaffected by the advertisement is not palatable to some researchers. On the other hand, the major advantage of the hypothesis test procedure over the confidence interval is its inherent simplicity, be it using the binomial probability distribution for an exact test result or the quick and easy normal approximation formula for an approximate test result.

Conclusions

The McNemar procedure is quick and easy to perform. The only assumption is that the outcome of each response is categorized into a dichotomy. For studies involving multichotomous categorical responses, an extension developed by the statistician Albert Bowker may be employed if the objective is to evaluate patterns of symmetry in the changes in response.

When evaluating the worth of a statistical procedure, famed statistician John Tukey defined "practical power" as the product of statistical power and the utility of the statistical technique. Based on this, the McNemar test enjoys a high level of practical power.

—Mark L. Berenson and Nicole B. Koppel

See also Inferential Statistics; Nonparametric Statistics; Paired Samples *t* Test (Dependent Samples *t* Test)

Further Reading

- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572–574.
- Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157.
- Tukey, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics*, 1, 31–48.

McNemar table/Excel spreadsheet: <http://blake.montclair.edu/~koppeln/mcnemar.htm> (Using Microsoft Excel's BINOMDIST function, exact *p* values of the McNemar test are obtained for all $n \leq 1,000$. A table of critical values for McNemar's test statistic *M*, along with their associated exact *p* values, was developed for all $n \leq 1,000$ using various traditional significance levels α for both one- and two-tailed tests. The corresponding spreadsheet can be downloaded from this site.)

MEAN

Of all the measures of central tendency, the mean is the most often used and can be defined in a variety of ways. It can be defined as the sum of all the scores in a data set divided by the number of observations, and can also be defined as the point about which the sum of the deviations is equal to zero.

The formula for the computation of the mean is as follows:

$$\bar{X} = \frac{\Sigma X}{n},$$

where

$\bar{X}$ (also called "X bar") is the mean value of the group of scores or the mean;

Σ (sigma) is the summation sign, which directs you to add together what follows it;

X is each individual score in the group of scores;

n is the size of the sample from which you are computing the mean.

Table 1 Sample Data for Computation of the Mean

Observation	Test Score 1	Test Score 2
1	7	14
2	8	13
3	6	15
4	7	21
5	5	31
6	6	27
7	4	28
8	7	21
9	6	32
10	8	25
11	9	23
12	7	24
13	8	21
14	9	18
15	7	19
16	6	25
17	6	22
18	7	23
19	4	27
20	5	31
21	6	21
22	5	25
23	4	29
24	8	34
25	7	20

For example, the data set in Table 1 consists of 25 cases with two variables, Test Score 1 and Test Score 2.

To compute the mean, follow these steps:

1. List the entire set of values in one or more columns such as you see in the table. These are all the X s.
2. Compute the sum or total of all the values.
3. Divide the total or sum by the number of values.

Applying the above formula to the sample data results in the following two means:

$$X_{Test1} = \frac{162}{25} = 6.48$$

$$X_{Test2} = \frac{589}{25} = 23.56$$

The mean is sometimes represented by the letter M and is also called the typical, average, or most central score.

More About the Mean

In the formula, a small n represents the sample size for which the mean is being computed. A large N represents the population size.

- The sample mean is the measure of central tendency that most accurately reflects the population mean.
- The mean is like the fulcrum on a seesaw. It's the centermost point where all the values on one side of the mean are equal in weight to all the values on the other side of the mean.
- The mean is very sensitive to extreme scores. An extreme score can pull the mean in one direction or another and make it less representative of the set of scores and less useful as a measure of central tendency.

Analysis Using SPSS

Figure 1 is a simple output using SPSS's descriptive feature.

—Neil J. Salkind

See also Measures of Central Tendency; Median; Mode

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

⇒ Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TestScore1	25	4	9	6.48	1.447
TestScore2	25	13	34	23.56	5.553
Valid N (listwise)	25				

Figure 1 Results of SPSS Descriptives Analysis

MEASUREMENT

It seems reasonable to expect that within the context of the *Encyclopedia of Measurement and Statistics*, one of the first questions that a reader might ask himself or herself would be, “What is the relationship between measurement and statistics?” Are these two terms synonymous, or is a distinction implied? Because the title of this encyclopedia uses both terms, perhaps we should attempt to understand the distinction between these two concepts. And perhaps it is for this reason that you have found yourself looking up the definition of measurement within this very volume.

Historical Perspectives on Measurement

For some, measurement is simply a tool of statistics, a necessary predecessor. We cannot perform statistical analyses if we have not measured anything. Consequently, some would say that measurement is simply the handmaiden of statistics. They might be inclined to take the point of view of S. S. Stevens, who defined measurement as “the assignment of numerals to objects or events according to a rule.” Stevens went on to propose four scales of measurement: nominal, ordinal, interval, and ratio.

Nominal scales are made up of variables with levels that are qualitatively different from one another. For example, consider the variable “ethnic background.” This variable could consist of any number of levels, but for the purposes of this example, let us consider just a few possibilities. A participant might report being Asian, African American, Latino, or Native American. Each of these different levels of the ethnic background variable represents a qualitatively different category. More specifically, each category is mutually exclusive of the others.

A second scale of measurement is an ordinal scale. Within the context of ordinal scales, the researcher is typically interested in knowing something about the rank order of the levels of the variable; however, there is no implication that these levels differ by any fixed

amount. Perhaps the classic example of ordinal-level measurement is found in the Olympic Games, where athletes receiving medals are ranked from 1st to 3rd place. It does not matter whether the gold medalist beat the silver medalist to the finish line by a fraction of a second or by 10 minutes. What matters is the order in which the athletes reached the finish line. The first one to cross the finish line receives a ranking of 1, and the second one across receives a ranking of 2. The assumption within an ordinal scale of measurement is that the fundamental quantity of interest is the relative placement and not the absolute difference between points on the scale. Another practical example of an ordinal scale of measurement would be the percentile ranking a student receives on an examination.

A third scale of measurement proposed by Stevens is the interval scale. Interval scales of measurement have several useful properties. For example, the difference between any two points on the scale is always the same. This is not always true within the context of an ordinal scale (e.g., percentile rankings). It is important to note, however, that interval scales do not have a true zero point—that is, a zero does not necessarily imply the complete absence of the construct of interest. Consider, for example, the measurement of intelligence. A person may receive a score of zero on a particular intelligence test if he or she answered all items incorrectly. Even if a person were to answer incorrectly every item on the test and receive a score of zero, this does not imply that the person has absolutely no intelligence. His or her intelligence was simply not captured by the items on that particular test (e.g., consider a Western intelligence test that is administered to participants from a very different culture).

By contrast, consider an example from the fourth scale of measurement, the ratio scale. Height is on a ratio scale, which is to say, there is a true zero point. Nothing exists that has a height of zero. A height of zero implies a total lack of the construct (i.e., absolutely no height). Other examples of ratio scales include temperature and weight. Besides retaining the desirable properties of the interval-level scale in which the difference between any points on the scale

is the same regardless of the placement on the continuum, the ratio-level scale has the desirable property of having a meaningful zero point. Most inferential statistical techniques assume that variables are measured at least on an interval-level scale, with the ideal being a ratio scale.

Modern Perspectives on Measurement

Although Stevens's definition of measurement and the corresponding four scales of measurement he set out continue to enjoy the limelight in the field of psychology as the popular conception of measurement, these ideas do not reign supreme. Modern theorists have suggested that measurement is not simply an operation that is performed on data in order to prepare them for statistical analysis. Rather, measurement and statistics themselves are fundamentally different concepts with different assumptions and different implications, as well as different rules about how best to approach analysis and interpretation.

Within such an alternative framework, the fundamental distinction between measurement and statistics is that in measurement, the researcher starts out with a predefined theoretical model of how the world looks. The researcher then collects some data and performs statistical tests to see whether the data fit the theoretical expectations of the theoretical model. If the data do not fit the model, then there are two options: (a) the data are bad, or (b) the theoretical model needs to be revised. By contrast, within the context of statistics, the researcher starts by gathering data related to a construct of interest and then attempts to use statistics to build a model that fits, or best explains, the patterns observed in the data.

Notice that these are radically different approaches to data analysis. Within the context of measurement, the theoretical models are falsifiable, which is a highly desirable property. Within the context of statistics, data analytic tools are used to help develop a theory about why the world (and the data) is (or are) a certain way. Michell outlines many of the limitations of psychological measurement in general when one

takes this latter atheoretical perspective. He argues that it is often assumed without evaluation that the measures used to develop and validate theories have quantitative structure. The implication is that when the measurement properties of the scores are unknown, the appropriateness of the conclusions is also unknown.

Unfortunately, the covert nature of the constructs in educational and psychological research means that it is very unlikely that a situation would exist where quantitative measurement can be assessed directly. In its stead, Michell advocates the work of Luce and Tukey on additive conjoint measurement as a means of assessing quantity indirectly. Conjoint measurement is concerned with the way the *ordering* of a dependent variable varies with the joint effect of two or more independent variables. For example, assume that C (complexity) represents the ordering of differences in task complexity, and that I (intelligence) represents the ordering of differences in intelligence (note that only ordering is assumed here). The dependent variable, P (performance), represents performance on an appropriate measure along which the effect of C and I is assessed. Therefore, the ordering of C and I is necessarily dependent upon the order of P. That is, their orders are relative to their effect on P, and if appropriate conditions hold, the variables C and I are quantified relative to their effects on P. Michell outlines the sufficient experimental conditions that satisfy conjoint measurement and therefore presents one approach for testing the presence of interval measurement.

The strongest advocates of the additivity concept of measurement within contemporary psychometrics are associated with the field of Rasch measurement. Wright, and Bond and Fox, have argued that there are at least two properties of the Rasch model central to fundamental measurement. The first is that the Rasch model tests the core assumption of conjoint additivity. The second property of the Rasch model essential to pure measurement is specific objectivity. Specific objectivity occurs when (a) estimates of the person's trait level (e.g., ability) are freed from the effects of the actual items

attempted, and (b) estimates of item difficulty are freed from the effects of the particular sample of people who answered the items. Sample-free estimates of test characteristics are important if one wants to generalize interpretations of measurements beyond the sample on which the test was developed. Specific objectivity is a function of the unique mathematical characteristics of the Rasch measurement model and is not present in other item response theory models.

In summary, Michell argues that fundamental measurement requires the presence of an additive structure in the data. Raw scores (i.e., numbers that are assigned to objects according to rules) may or may not represent interval (additive) levels of measurement, and the assumption that they do must be tested. Bond and Fox argue that most often, by simply following the assignment of numbers to objects according to rules—the dominant but naïve view of measurement inculcated by Stevens—the researcher is left with a score that is at best measured at the ordinal level. Unfortunately, researchers almost invariably then treat this score as if it represented interval-level measurement (a typical requirement for subsequent statistical analyses) without recognizing that there is a way to empirically test the assumption of additivity via the Rasch measurement model (or using the procedures advocated by Michell). Assuming that variables are measured at an interval level and summing them to create a composite score will potentially lead to invalid and misleading results. That is, interpretation of the statistics applied to the analysis of test scores, no matter how carefully constructed the test is, will be qualified to the extent that the scores deviate from an interval scale measure. Rather than simply assuming additivity, we are obliged to test the limits of this assumption before applying our statistics.

Conclusion

Measurement and statistics represent fundamentally different approaches to analysis. Measurement is concerned with examining the extent to which data fit a model, and statistics is concerned with building a

model to fit the data. But the concepts of measurement and statistics do have a symbiotic relationship. At best, when fundamental measurement is achieved, the researcher is able to develop a set of variables that meets the strict assumptions of her or his preferred theoretical model. When this kind of pure measurement is achieved, the researcher is then in a position to use various statistical techniques to analyze the data and will be in a strong position to contribute valid, replicable findings to the research community.

—Steven E. Stemler and Damian P. Birney

See also Reliability Theory; Validity Theory

Further Reading

- Bond, T., & Fox, C. (2001). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–384.
- Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, 10(5), 639–667.
- Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237–255.
- Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*. New York: Wiley.
- Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.

Rasch Measurement Special Interest Group of the American Educational Research Association: <http://www.raschsig.org>; see also <http://www.rasch.org/rmt/index.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Essex, C., & Smythe, W. E. (1999). Between numbers and notions: A critique of psychological measurement. *Theory & Psychology*, 9(6), 739–767.

When certain mathematical machinery is applied to ideas from psychology or education, that machinery imposes certain requirements in the relationship between numbers and notions. The result can be a theory-neutral approach to research, where commitments in response to the options are made unknowingly, thus becoming instead a theory-by-default endeavor. This paper catalogs some of these mathematical choices to help make theoretical commitments more explicit.

MEASUREMENT ERROR

Measurement error can be described as the variability in measurements of the same quantity on the same item. These errors occur for a variety of reasons, including inadequate survey design, sampling variability, inherent biological variation, and laboratory error analysis. Measurement error is often classified into two broad categories: random error and systematic error.

Random error refers to any random influence on the measurement of a variable. Consider, for example, a person whose weight is taken at a hospital. If two different nurses see the patient, they may not record the same weight, or if the same nurse takes two measurements only minutes apart, different eye positions may cause different scale readings. There is no discernible pattern, so the measurement may be higher or lower than the true value. Therefore, the sum of all random error in a series of measurements on the same variable should, in theory, equal zero. Thus, random error does not strongly influence the average value of the measurement. This type of error is sometimes referred to as noise.

Systematic error causes measurements to be consistently higher or lower. For example, suppose that the scale used to weigh patients consistently

shows readings that are five pounds heavier than the true value. This suggests that the scale should be recalibrated. Systematic errors can be minimized or completely eliminated by the use of careful planning in an experiment. This type of error is commonly referred to as bias.

The statistical models and methods for analyzing data measured with error are called measurement error models. A typical problem in statistics is to determine the linear relationship between two quantitative variables, say, X and Y , in order to use X to explain or predict Y . In this case, we assume that X cannot be observed, and a substitute variable W is used. Often, we write $W = X + U$, where U is the measurement error. One problem of interest is how the linear relationship between W and Y differs from the relationship between X and Y . It can be shown that the best-fitting line that relates W to Y is biased toward zero. Thus, using W as a substitute restricts one's ability to accurately assess the true linear relationship between X and Y . Various techniques have been developed to correct for such problems.

It is difficult to completely avoid measurement error; however, one can take several steps to reduce it. First, make sure that any instruments used are tested initially, monitored over time, and recalibrated as needed. Second, if possible, use only one instrument or set of instruments to do all of the measuring in an experiment. Third, use statistical methods that account for the presence of measurement error when analyzing data. Finally, take repeated measurements on the same variable.

—Kimberly Weems

See also Measurement; Reliability Theory

Further Reading

- Buzas, J., Tosteson, T., & Stefanski, L. (2003). *Measurement error*. Institute of Statistics Mimeo Series No. 2544. Raleigh: North Carolina State University.
- Trochim, W. (2002). *Measurement error*. Retrieved from <https://www.socialresearchmethods.net>

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are measures of the location of the center or middle of a distribution. However, the definition of “center” or “middle” is deliberately left broad, such that the term *central tendency* can refer to a wide variety of measures. The three most common measures of central tendency are the mode, the mean, and the median.

Mode

The mode for a collection of data values is the data value that occurs most frequently (if there is one). Suppose the average number of colds in a family of six in a calendar year is as presented in Table 1.

Then, the mode is 1 because more family members (i.e., $n = 2$) caught one cold than any other number of colds. Thus, 1 is the most frequently occurring value. If two values occur the same number of times and more often than the others, then the data set is said to be bimodal. The data set is multimodal if there are more than two values that occur with the same greatest frequency. The mode is applicable to qualitative as well as quantitative data.

With continuous data, such as the time patients spend waiting at a particular doctor’s office, which can be measured to many decimals, the frequency of each value is most commonly 1 because no two scores will be identical. Consequently, for continuous data, the mode typically is computed from a grouped frequency distribution. The grouped frequency distribution in Table 2 shows a grouped frequency distribution for the waiting times of 20 patients. Because the

Table 1 Number of Colds in a Selected Family

<i>Family Member</i>	<i>Frequency</i>
Father	5
Mother	4
First Son	1
Second Son	2
First Daughter	1
Second Daughter	3

Table 2 Grouped Frequency Distribution

<i>Range</i>	<i>Frequency</i>
0 – <10	2
10 – <20	2
20 – <30	3
30 – <40	7
40 – <50	3
50 – <60	2
60 – <70	1

interval with the highest frequency is 30 – <40 minutes, the mode is the middle of that interval (i.e., 35 minutes).

Mean

Arithmetic Mean

The *arithmetic mean*, or average, is the most common measure of central tendency. Given a collection of data values, the mean of these data is simply the arithmetic average of these data values. That is, the mean is the sum of observations divided by the number of observations. If we use the following notation:

x is the variable for which we have data (e.g., test scores),

n is the number of sample observations (sample size),

x_1 is the first sample observation (first test score),

x_2 is the second sample observation (second test score),

x_n is the n th (last) sample observation (last test score),

then the sample mean of a sample $x_1, x_2, \dots, x_n$ is denoted by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}.$$

To find the average number of colds in the family presented earlier, we compute

$$\bar{x} = \frac{5 + 4 + 1 + 2 + 1 + 3}{6} = \frac{16}{6} = 2.67.$$

The arithmetic mean is not the only “mean” available. Indeed, there is another kind of mean that is called the geometric mean, which is explained below. However, the arithmetic mean is by far the most commonly used. Consequently, when the term *mean* is used, one can assume that it is the arithmetic mean.

Weighted Mean

The weighted mean of a set of measurements $x_1, x_2, \dots, x_n$ with relative weights $w_1, w_2, \dots, w_n$ is given by

$$\bar{x} = \frac{\sum xw}{\sum w}.$$

The weighted mean has many applications. In effect, it is used to approximate the mean of data grouped in a frequency distribution. In order to approximate the mean waiting time of the 20 patients presented above, the class mark is used to represent the waiting time of each person falling within that class. A weighted mean is then calculated, where the x s are the class marks and the weights are the corresponding class frequencies, as in Table 3.

$$\begin{aligned}\bar{x} &= \frac{\sum xw}{\sum w} \\ &= \frac{670}{20} = 33.5\end{aligned}$$

Thus, the weighted mean is 33.5 minutes. The mean has two important properties. First, the sum of the deviations of all scores in the distribution from the mean is zero. Second, the sum of squares of deviations about the mean is smaller than the sum of

squares of deviations about any other value. Consequently, the mean is the measure of central tendency in the *least squares* sense inasmuch as the sum of the squared deviations is a minimum.

Trimean

The trimean is another measure of central tendency. It is computed by adding the 25th percentile plus twice the 50th percentile plus the 75th percentile, and then dividing by four. The 25th, 50th, and 75th percentile of the cold data set is 1, 2.5, and 4.25, respectively. Therefore, the trimean is computed as

$$\frac{1 + (2 \times 2.5) + 4.25}{4} = 2.56.$$

The trimean value of 2.56 is close to the arithmetic mean value of 2.67. The trimean has logical appeal as a measure of central tendency. However, it is rarely used.

Trimmed Mean

The trimmed mean is computed by discarding a certain percentage of the lowest and the highest scores in a ranked (i.e., ordered) set of data and then computing the mean of the remaining scores. For example, a mean trimmed 50% is computed by discarding the highest and lowest 25% of the scores and taking the mean of the remaining scores. The mean trimmed 0% provides the arithmetic mean. Trimmed means are used in certain sporting events (e.g., ice skating, gymnastics) to judge competitors' levels of performance and to prevent the effects of extreme ratings possibly caused by biased judges. Before scores are discarded, the analyst must first rank the data. For the cold data, the mean trimmed 33% would result in the highest value (i.e., 5) and lowest value (i.e., 1) being discarded, resulting in the following trimmed mean:

$$\bar{x} = \frac{1 + 2 + 3 + 4}{4} = \frac{10}{4} = 2.5.$$

Geometric Mean

The geometric mean of n numbers is obtained by multiplying all of them together, and then taking the

Table 3 Computation of Weighted Mean

Waiting Time	Class Mark (x)	Frequency (w)	x·w
0 – <10	5	2	10
10 – <20	15	2	30
20 – <30	25	3	75
30 – <40	35	7	245
40 – <50	45	3	135
50 – <60	55	2	110
60 – <70	65	1	65

n th root of them. In other words, the geometric mean is the n th root of the product of the n scores in the dataset. Thus, the geometric mean of the cold data—5, 4, 1, 2, 1, and 3—is the sixth root of $5 \times 4 \times 1 \times 2 \times 1 \times 3$, which is the sixth root of 120 (because there are six numbers), which equals 2.22. The formula can be written as

$$\text{Geometric mean} = (\Pi X)^{1/n},$$

where ΠX means to take the product of all the values of X , and the superscript value (i.e., $1/n$) indicates the n th root. The geometric mean can also be computed by

1. computing the logarithm of each number,
2. computing the arithmetic mean of the logarithms,
3. raising the base used to take the logarithms to the arithmetic mean.

Thus, if the natural logarithm (i.e., Ln) is used, then raising this base would necessitate use of the exponent. For the cold data, the computation would be as in Table 4.

The base of natural logarithms is 2.718. The expression $\text{EXP}[0.7979]$ means that 2.718 is raised to the 0.7979th power. $\text{Ln}(X)$ is the natural log of X .

An identical result can be obtained by using logs base 10 as shown in Table 5.

If any one of the scores is zero, then the geometric mean is zero. If any scores are negative, then the geometric mean is meaningless. The geometric mean

Table 4 Computation of Geometric Mean Using Natural Logarithms

<i>Family Members</i>	<i>Frequency (x)</i>	<i>Ln(X)</i>
Father	5	1.6094
Mother	4	1.3863
First Son	1	0.0000
Second Son	2	0.6931
First Daughter	1	0.0000
Second Daughter	3	1.0986
	Arithmetic mean	0.7979
	Exponential	$\text{EXP}[0.7979] = 2.22$
	Geometric mean	2.22

Table 5 Computation of Geometric Mean Using Base 10 Logarithms

<i>Family Members</i>	<i>Frequency (x)</i>	<i>Ln(X)</i>
Father	5	0.6990
Mother	4	0.6021
First Son	1	0.0000
Second Son	2	0.3010
First Daughter	1	0.0000
Second Daughter	3	0.4771
	Arithmetic mean	0.3465
	Exponential	$100.3465 = 2.22$
	Geometric mean	2.22

is an appropriate measure to use for averaging rates. However, it is one of the least used measures of central tendency.

Harmonic Mean

The harmonic mean is the mean of n numbers expressed as the reciprocal of the arithmetic mean of the reciprocals of the numbers. The harmonic mean typically is used to take the mean of sample sizes. For the cold data, the harmonic mean is defined as

$$\bar{x}_h = \frac{i}{\frac{1}{n_1} + \frac{1}{n_2} + \dots + \frac{1}{n_i}},$$

where i is the number of scores from which the harmonic mean is computed. For the cold data, the harmonic mean is

$$\bar{x}_h = \frac{6}{\frac{1}{5} + \frac{1}{4} + \frac{1}{1} + \frac{1}{2} + \frac{1}{1} + \frac{1}{3}} = 1.83.$$

This is less than the arithmetic mean of 2.67, the trimean of 2.56, and the geometric mean of 2.22.

Median

The median is the midpoint of a distribution such that the same number of scores is above the median as below it. In other words, the median is the 50th percentile. More specifically, the median for a collection of data values is the number that is exactly

in the middle position of the list when the data are ranked (i.e., arranged in increasing order of magnitude). The formula for the median (Md) is

$$Md = L + \left(\frac{N(0.50) - cfb}{fw} \right) (i)$$

where

L is the lower limit of the interval within which the median lies,

N is the number of cases in the distribution,

cfb is the cumulative frequency in all intervals below the interval containing the median,

fw is the frequency of cases within the interval containing the median,

i is the interval size.

In order to compute the median for the cold data, the first step is to rank the data:

1, 1, 2, 3, 4, 5.

Because there are six numbers (i.e., an even number of data points), there are two middle numbers, namely, 2 and 3. Therefore, $L = 1.5$ (i.e., the lowest of the two middle numbers – 0.5). Also, $N = 6$ (number of observations), and $cfb = 2$ (i.e., the number of observations that lie below the lower limit of 1.5). Also, $fw = 2$ (i.e., the number of observations that are equal to the middle numbers) because the two middle numbers (i.e., 2 and 3) do not appear anywhere else in the data set. Finally, $i = 2$ (i.e., the highest middle number – the lowest middle number + 1 = 3 – 2 + 1 = 2). (Please note that $i = 1$ if the middle numbers are all the same.) Thus, the median is

$$Md = 1.5 + \left(\frac{6(0.50) - 2}{2} \right) (2) = 2.5.$$

Thus, the median cold is 2.5. When the number of observations is relatively small and the data are not grouped in class intervals—as is the case with the cold data—the median can be computed using the following steps:

1. Order the n observations from smallest to largest, including any repeated observations, so that every observation appears in the list.
2. Determine the location of the sample median, which is given by $(n + 1)/2$. Thus, for example, for a sample size of 5 (i.e., $n = 5$), $(n + 1)/2 = 3$, and the median is represented by the third number in the series. For a sample size of 6, $(n + 1)/2 = 3.5$, and the median can be located somewhere between the third and fourth number in the series.
3. If the number of scores is odd, the median is the middle score. Consider the following ranked distribution of scores: 1, 3, 3, 5, 6, 7, 8, 8, 9. Because there are nine scores (i.e., $N = 9$), the median is 6.
4. If the number of scores is even, the median is the average of the two middle values. Thus, because the cold data have an even number of scores, the two middle numbers are 2 and 3, and the median is the average of 2 and 3 (i.e., the average of the third and fourth observations), which is 2.5.

It can be seen that the simpler method of calculating the median yielded exactly the same number as did using the more general formula. However, for relatively large sample sizes, the simpler formula can distort the true value of the median represented by the more general formula.

Computer Applications

When using SPSS, there are a few ways to compute measures of central tendency. The Frequencies command can be used to compute the mean, median, and mode. The Descriptives command can be used to compute the mean. The Explore command can be used to compute the median, mean, and trimmed mean. The Means command can be used to compute the median, mean, harmonic mean, and geometric mean. Finally, the Reports command can be used to compute the mean and median.

The SPSS output for the Frequencies command pertaining to the cold data set is presented in Figure 1.

The SPSS output for the Descriptives command pertaining to the cold data set is in Figure 2.

The SPSS output for the Explore command pertaining to the cold data set is in Figure 3.

COLDS

N	Valid	6
	Missing	0
Mean		2.67
Median		2.50
Mode		1

Figure 1 Measures of Central Tendency Using Frequencies Command

	N	Mean
COLDS	6	2.67
Valid N (listwise)	6	

Figure 2 Measures of Central Tendency Using Descriptives Command

The SPSS output for the Means command pertaining to the cold data set is in Figure 4.

The SPSS output for the Reports command pertaining to the cold data set is in Figure 5.

Comparisons of Measures of Central Tendency

To some extent, selection of the most appropriate measure of central tendency is dependent on the scale of measurement of the variable. Specifically, if the

	Statistic	Std. Error
COLDS Mean	2.67	.67
95% Confidence Interval for Mean	.95	
Lower Bound	4.38	
Upper Bound	2.63	
5% Trimmed Mean	2.50	
Median	2.667	
Variance	1.63	
Std. Deviation	1	
Minimum	5	
Maximum	4	
Range	3.25	
Interquartile Range	.383	.845
Skewness	-1.481	1.741
Kurtosis		

Figure 3 Measures of Central Tendency Using Explore Command

COLDS

SAMPLE	Mean	Median	Harmonic Mean	Geometric Mean
1	2.67	2.50	1.83	2.22
Total	2.67	2.50	1.83	2.22

Figure 4 Measures of Central Tendency Using Means Command

data are nominal, then only the mode is appropriate. If the data are ordinal, either the mode or the median may be appropriate. If the data are interval or ratio, the mode, median, or mean may be appropriate.

For distributions that are symmetrical and unimodal, the three major measures of central tendency (i.e., mean, median, mode) are all the same. When the distribution is symmetrical and bimodal, the mean and the median coincide, but two modes are present. The less symmetrical the distribution, the greater the differential between the mean, the median, and the mode. For skewed distributions, they can differ markedly. Specifically, in positively skewed distributions, the mean is higher than the median, whereas in negatively skewed distributions, the mean is lower than the median. Thus, comparing the mean and median can provide useful information about the level of skewness inherent in the distribution.

Of the eight measures of central tendency discussed, the mean is by far the most widely used because it takes every score into account, is the most efficient measure of central tendency for approximately symmetric (normal) distributions, and uses a simple formula. Also, because the mean requires that the differences between the various levels of the categories on any part of the distribution represent equal differences in the characteristic or trait measured (i.e., equal unit or interval/ratio scale), it can be manipulated mathematically in ways not appropriate to the median and mode. Thus, the mean is mathematically appealing, making it possible for

COLDS

N	Mean	Median	Harmonic Mean	Geometric Mean	Grouped Median
6	2.67	2.50	1.83	2.22	2.50

Figure 5 Measures of Central Tendency Using Reports Command

researchers to develop statistical procedures for drawing inferences about means. However, the mean does have several disadvantages. In particular, the mean is sensitive to skewed data. It is also sensitive to outliers. Thus, the mean often is misleading in highly skewed distributions and is less efficient than other measures of central tendency when extreme scores are possible.

The trimean is almost as resistant to extreme scores as is the median, and it is less subject to sampling fluctuations than the arithmetic mean in extremely skewed distributions. However, it is less efficient than the mean for normal distributions. The trimmed mean, which generally falls between the *mean* and the *median*, is less susceptible to the effects of extreme scores than the arithmetic mean and, in turn, is less susceptible to sampling fluctuation than the mean for extremely skewed distributions. However, like the trimean, the trimmed mean is less efficient than the mean for normal distributions. The geometric mean is less affected by extreme values than the arithmetic mean and is useful as a measure of central tendency for some positively skewed distributions. However, the geometric mean is rarely used because (a) it equals zero if any one of the scores is zero, regardless of how large the remaining scores are; (b) it is meaningless if any scores are negative; and (c) it is more difficult to compute than the arithmetic mean. The weighted mean does not use any of the actual scores in the distribution.

The median is useful because of its ease of interpretation and because it is more efficient than the mean in highly skewed distributions. That is, the median is not sensitive to skewed data. However, it does not take into account every score, relying only on the middle value(s) in an ordered set of data. Also, the median

generally is less efficient than the mean, the trimean, and the trimmed mean. The mode can be informative, is easy to interpret, and is the only measure of central tendency that can be used with nominal data; however, it should almost never be used as the only measure of central tendency because it depends only on the most frequent observation and is

highly susceptible to sampling fluctuations. Another disadvantage of the mode is that many distributions have more than one mode, thereby complicating interpretation. Also, the mode does not always exist.

—Anthony J. Onwuegbuzie,
Larry Daniel, and Nancy L. Leech

See also Mean; Median; Mode

Further Reading

- Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
- Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). *SPSS for basic statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.

Means and medians graphical comparison applets: http://www.ruf.rice.edu/~lane/stat_sim/descriptive/ and <http://standards.nctm.org/document/examples/chap6/6.6/>

MEDIAN

The median is a measure of central tendency and is the point in a group of values with an equal number of values above and below that point.

The computation of the median is as follows.

- For an odd number of values, the position of the median is given by $(N + 1)/2$. If we have 15 cases, the median is the 8th case.
- For an even number of cases $(N + 1)/2$ does not give a whole number. In this case, the median is the (arithmetic) mean of the two values immediately above and below.

For example, the data set in Table 1 consists of 11 test scores.

First of all, the values need to be sorted from smallest to largest value: 4, 5, 6, 6, 6, 7, 7, 7, 8, 8, 9. Then, the midpoint needs to be found that splits the sample into two halves. In this example we have an odd number of participants, so we will have a midpoint that represents the point on which an equal number of values on each side is present. To the left of the midpoint, the values 4, 5, 6, 6, and 6 are present, and to the right, the values 7, 7, 8, 8, and 9. In this example, the midpoint of the row of numbers has a value of 7, so the median is 7.

Consider also a sample of an even number of participants who have achieved the test scores 4, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, and 9. The midpoint has to be computed by finding the two values in the middle—here, 6 and 7—and the mean of the two will be the median, so the median is 6.5.

Table 1 Sample Data Set

Participant	Test Score
1	7
2	8
3	6
4	7
5	5
6	6
7	4
8	7
9	6
10	8
11	9

More About the Median

- In a group with an odd number of values, the median is a specific value (an identifiable individual) in the group. It is not computed by directly taking the other values of the group into account. In a group with an even number of values, the mean of the two middle values is taken. Other than these two, no other values enter the equation.

Frequencies

Statistics		
testscore		
N	Valid	12
	Missing	0
Median		6.50

Frequencies

Statistics		
testscore		
N	Valid	11
	Missing	0
Median		7.00

Figure 1 SPSS Output

- The median is a midpoint that splits a group into two halves equal in their number of values but not equal in their actual values.
- The median is not sensitive to extreme scores. The magnitude of the other values in the group and their relative magnitude in comparison to the median are not taken into account. In a group with the values 1, 2, and 99, 2 is the median. An extreme score can pull the mean in one or another direction and make it less representative of the set of scores and less useful as a measure of central tendency.
- The median can be a more useful descriptor in a skewed distribution than the mean.

Analysis Using SPSS

Figure 1 shows a simple output using SPSS's descriptive feature.

—*Susanne Hempel*

See also Average; Estimates of the Population Mean; Mean; Measures of Central Tendency; Mode

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

MEDIAN TEST

Many statistical procedures test for differences in location parameter, where the measure of location pertains to the median (θ), or the 50th percentile. Hypotheses can be tested about a single or multiple medians.

$k = 1$ Median

Testing hypotheses about a single sample median presents two problems. First, and perhaps more debilitating, the sample median is not the best unbiased estimate of the population median. (Therefore, the question arises as to the efficacy of testing this hypothesis.)

The Harrell-Davis (θ_{HD}) is one statistic, among many, that is a better estimate of the population median than the sample median. The θ_{HD} is computed by drawing N ordered deviates from the beta distribution. Sort the original scores (x_i), and multiply the first score by the first beta deviate (b_1), the second score by the second beta deviate, and so forth. The Harrell-Davis estimator of the population median is

$$\theta_{HD} = \sum_{i=1}^N b_i x_i. \quad (1)$$

The second problem is that the sampling distribution is intractable. Thus, it is only possible to estimate the standard error of θ . One suitable estimate of the standard error is the Maritz-Jarrett, s , which in turn may be based on the Harrell-Davis estimator of the median.

Thus, for $k = 1$, the null hypothesis is $H_0: \theta = \theta_0$, where θ is the sample median and θ_0 is some hypothesized population value. This is tested against the alternative $H_1: \theta \neq \theta_0$. The test is

$$Z = \frac{\theta - \theta_0}{s}, \quad (2)$$

where s is the Maritz-Jarrett estimate. Wilcox provided algorithms and computer code to automate computing the Maritz-Jarrett estimate. The significance of Z for a specified α level is determined by a table of the standard normal distribution.

$k > 1$ Medians

A classical test for a difference in two or more ($k \geq 2$) independent medians is due to Mood. The two-sample case is analogous to Fisher's exact test and the $k > 2$ is based on the chi-square, and they share the assumptions of the analogous tests.

Consider the case of $k = 3$. The null hypothesis $H_0: \theta_A = \theta_B = \theta_C$, which is tested against the alternative hypothesis H_a : at least one θ is different from the rest. The grand median is computed. The number of scores above and below the median is tabulated. The expected value for each group is half the sample size for that group. The chi-square test is computed, and significance is determined by comparing the obtained chi-square statistic with the critical value based on $df = k - 1$.

Example

Three groups of test scores are presented in Table 1. The grand median is 65.5 (Fay has provided advice concerning values tied with the grand median).

Count the number of scores greater than 65.5 and the number of scores less than or equal to 65.5 for each group. The frequencies are displayed in Table 1.

Table 1 Test Scores for Groups 1, 2, and 3

Group 1	Group 2	Group 3
63	69	35
99	73	38
89	47	61
44	33	84
88	63	74
70	68	26
66	70	49
84	53	89
93	83	68
66	73	66
96	40	40
37	58	36
46	82	32
78	48	73
95	25	50
49	92	52
61	64	30
97	42	70
95	71	39
75	75	69
65	72	37
70	34	62
44	28	87
67	64	44
48	76	89
54	31	84

Table 2 Frequencies Above and Below the Grand Median

	Group 1	Group 2	Group 3
>65.5	16	12	11
<65.5	10	14	15

The expected value for each cell is 13. This was found by dividing the sample size for each group ($n_1 = n_2 = n_3 = 26$). The values for each of the six cells are calculated based on

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (3)$$

where O is the observed value, E is the expected value, i is the row, j is the column, and k is the number of groups. The 16 participants in Group 1 had a score above 65.5. Thus,

$$\frac{(16 - 13)^2}{13} = \frac{(3)^2}{13} = \frac{9}{13} = 0.692.$$

χ^2 = the sum for the six cells, as indicated in Table 3.

χ^2 is 2.154, and $df = (k - 1) = 3 - 1 = 2$. The critical value for χ^2 is 5.991. Because $2.154 < 5.991$, the decision is to fail to reject the null hypothesis that there is no difference in the medians of the three groups.

Calculation Using SPSS

The data must be entered into one variable—in this case, “scores”—using a grouping variable. Table 4 shows how to use the grouping variable for the first few data points.

To conduct Mood’s test (called the median test in SPSS), select the following: Analyze | Nonparametric

Table 3 Results of the Calculations for Each Cell

	Group 1	Group 2	Group 3	Total
>65.5	0.692	0.077	0.308	1.077
≤65.5	0.692	0.077	0.308	1.077
Total	1.385	0.154	0.615	2.154

Table 4 Using a Grouping Variable in SPSS

Group	Score
1	67
1	48
1	54
2	69
2	73
2	47
3	35
3	38
3	61

Tests | K Independent Samples. Then, enter the dependent variable into the area designated as “Test Variable List.” Enter the grouping variable into the area “Grouping Variable.” Click the button “Define Range” and enter the minimum and maximum numbers for the grouping variables. In the rectangle titled “Test Type,” select the median test using the checkbox to the left. The SPSS output for the median test includes the frequency table and test statistic, as indicated in Tables 5 and 6.

Table 5 Frequencies

		group		
		1	2	3
score	> Median	16	12	11
	≤ Median	10	14	15

Table 6 Test Statistics

	Score
N	78
Median	65.50
Chi-Square	2.154 ^a
df	2
Asymp. Sig.	.341

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 13.0.

See also Estimates of the Population Median; Inferential Statistics; Nonparametric Statistics

Further Reading

- Fay, B. R. (2006). The effect Type I error and power of various methods of resolving ties for six distribution-free tests of location. *Journal of Modern Applied Statistical Methods*, 5(1), 50–67.
- Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73, 194–196.
- Mood, A. M. (1950). *Introduction to the theory of statistics*. New York: McGraw-Hill.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

META-ANALYSIS

Meta-analysis is a method of analysis that allows for the statistical aggregation of data from individual empirical findings. This is conducted by transforming results from a pool of studies into one common metric, the effect size (ES). The earliest form of research synthesis was performed as early as 1904 by biometrician Karl Pearson in order to review the empirical basis for the routine use of the typhoid vaccine in the British army for soldiers at risk for the disease. Pearson estimated inoculation effectiveness by averaging correlations reported in five studies about the relationship between inoculation status and typhoid immunity as well as those reported in six studies about the relationship between inoculation status and mortality among those who suffered from the condition. The modern usage of this statistical procedure and the coining of the term *meta-analysis* began with the work of Gene Glass in 1976. Applications of the method by individuals such as Mary L. Smith and Gene Glass to psychotherapy, John E. Hunter and Frank L. Schmidt to validity of employment tests, and Robert Rosenthal and Donald B. Rubin to interpersonal expectancy effects highlighted the utility of the meta-analysis in the social sciences. In 1985, Larry V. Hedges and Ingram Olkin, with the publication of *Statistical Methods for*

Meta-Analysis, helped to introduce the method as a statistical specialty.

Stages of Research Synthesis in a Meta-Analysis

Conducting a systematic review involves the aggregation and comparison of findings across empirical research studies presenting quantitative data. Although this continues to be an important issue of debate in the field of meta-analysis, many experts agree that a useful research synthesis should be based on findings from high-quality studies with methodological rigor. Relaxed inclusion standards for studies in a meta-analysis may lead to a problem that Hans J. Eysenck in 1978 labeled as “garbage in, garbage out.” Several technical objections have been noted by critics of meta-analysis such as John S. Searles: (a) The meta-analysis approach mixes apples and oranges in combining the findings of studies with varying methodological quality; (b) the meta-analysis approach aggregates the findings of poor studies, thus setting low standards of judgment for quality of outcome study; (c) the meta-analysis approach is problematic in light of shortcomings and flaws in the published literature (e.g., selective reporting, bias, insufficient reporting of primary data); (d) the meta-analysis approach has used “lumpy,” nonindependent data (i.e., multiple uses of the same data in several studies); and (e) the meta-analysis approach has given equal weighting to all studies regardless of methodological rigor.

According to Harris Cooper and Larry V. Hedges, a methodologically robust meta-analysis needs to be completed in five organized stages. The first stage of research synthesis in meta-analysis is to formulate an appropriate research hypothesis. For instance, one may question whether a specific independent variable causes systematic changes in specific dependent or outcome variables. Other issues of interest may be the generalizability or the external validity of findings from specific populations, settings, and procedures to a larger set of studies. Answers to such research questions can then advance the theoretical understanding of a specific phenomenon in a meaningful manner.

The second stage of research synthesis focuses on identification of studies and subsequent data collection from these primary sources. Clear eligibility guidelines in terms of inclusion and exclusion criteria will be helpful in identifying appropriate studies. Several methods may be used to locate and retrieve empirical reports. One strategy is to perform keyword searches of selected computerized bibliographic databases such as *Psychological Abstracts* (PsycINFO or PsychLit), *Sociological Abstracts*, ERIC (Educational Resources Information Center), SSCI (Social Science Citation Index), or MEDLINE. Computer searches, however, have been criticized for their unreliability and inaccuracy in locating relevant studies in meta-analyses. Additional methods may be employed for retrieval of appropriate studies: (a) manual search of journals that routinely publish research relevant to the topic selected; and (b) usage of the ancestry approach (i.e., “footnote chasing”) where the reference sections of relevant literature are reviewed for eligible studies. In order to effectively minimize publication bias or the “file drawer” effect (i.e., low publication probability for nonsignificant research findings), a comprehensive search of fugitive literature may also be conducted. For instance, one may search for master’s theses and dissertations by using the Dissertation Abstracts International (ProQuest Digital Dissertations) database. Obvious advantages in using this method are the availability of theses and dissertations as well as their detailed description of methodology and results.

Data collection, through the process of coding, is one of the most demanding tasks involved in the conduct of a meta-analysis. In order to extract information from each eligible study, a coding manual may be developed by the meta-analyst(s). A coding protocol is conceptualized as consisting of two major parts, with the first part designed to collect data about the factors that may potentially influence the nature and value of the overall effect size (i.e., study descriptors) and the second part designed to collect information about specific study findings (i.e., effect size values for dependent variables). Because each study is expected to report varying numbers and forms of effect size statistics,

the coding form may be constructed in such a way as to offer the coder the flexibility of recording statistical information in a number of different formats. Reliability checks of the coding can then be performed during the continuous training of study coders. In cases of meta-analyses performed by a single analyst, Mark W. Lipsey and David B. Wilson recommend that coder reliability could be verified by selecting a subsample of coded studies and coding them again after sufficient time has passed since the first coding. Coder reliability (i.e., agreement rate) can then be obtained by dividing the number of observations agreed upon by the total number of observations, as suggested by Robert G. Orwin.

The third stage of research synthesis involves a careful evaluation of the data collected. A comprehensive review of these evaluation strategies is beyond the purpose of this entry, and the reader should refer to the reading list provided. The fourth stage is analysis and interpretation (see section below on statistical methods of effect size estimation). The fifth stage is the public presentation of the findings.

Statistical Methods of Effect Size Estimation

Two types of effect sizes have been described by Robert Rosenthal: the r family and the d family. The r family consists of the Pearson product moment correlation, whereas the d family includes Hedges’s g , Glass’s Δ , and Cohen’s d . An effect size for each study is calculated, and the mean effect size is then obtained from averaging all effect sizes calculated for each study. A widely used convention of evaluating the magnitude of an effect size is based on Jacob Cohen’s d estimates of small, moderate, and large effect sizes as 0.20, 0.50, and 0.80, respectively. Once effect sizes are calculated from a number of different studies, then particular attention is given to detecting any variability across these findings. In cases of observed variability, specific moderators that might have had an impact on effect sizes are consequently examined.

Brief Example of Effect Size Estimation

Effect sizes may be calculated based on the standardized mean difference index (ES_{sm}) or Cohen's d , which is used to synthesize data from studies that contrast two independent groups on measures that have a continuous distribution. A formula for ES_{sm} is the difference between the posttreatment group means divided by the posttreatment pooled standard deviation,

$$ES_{sm} = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{S_{pooled}}, \quad (1)$$

where $\bar{X}_{G1}$ is the posttreatment Group 1 mean, $\bar{X}_{G2}$ is the posttreatment Group 2 mean, and S_{pooled} is the posttreatment pooled standard deviation, which is estimated by Equation 2 presented below:

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \quad (2)$$

where n_1 and n_2 are posttreatment sample sizes (i.e., completers) for Group 1 and Group 2, respectively; standard deviations for Groups 1 and 2 are represented by s_1 and s_2 .

Cohen's d is a biased estimate of the population d effect size and needs to be corrected. In particular, for this effect size index, the upward bias may become non-negligible when sample sizes are less than about 20. Larry V. Hedges's and Ingram Olkin's d , which is an unbiased estimate, may be calculated by the investigator for correcting this bias.

Methods of Identifying and Correcting Publication Bias

Publication bias can have rather deleterious effects on the validity of a meta-analysis and consequent inferences drawn from its results. According to Colin Begg, publication bias occurs when a study is selected to be published because it has estimated the largest effect size. This, however, provides an overly optimistic estimate of the "true (mean) effect," whose magnitude is a function of the sample size of the study as well as the number of studies from which the effect size is estimated. Robert Rosenthal describes this

failure of journal editors to publish or the failure of original investigators to report their nonsignificant findings as a censorship of effect size estimates. Unpublished nonsignificant results are placed in the "file drawers" of the researchers.

A possible strategy to detect potential publication bias is to obtain a "funnel plot." A funnel plot is a scatterplot of sample size versus estimated effect size for all included studies in a meta-analysis. If the plot obtained does not resemble the shape of a funnel, then the possibility of a publication bias is considered highly likely. Some meta-analysts comment that funnel plots may be difficult to interpret. Another method of detecting publication bias involves an examination of the correlation between sample size and effect size. A Fail-Safe-N (FSN) may also be calculated. According to Robert Rosenthal, this approach calculates the number of studies with an effect size of zero needed to reduce the mean obtained effect size to the point of nonsignificance.

Ruling Out Other Explanations for Significant Effect Sizes

Joseph A. Durlak has recommended that prior to making interpretations about effect sizes, other explanations that may have accounted for positive effects should be analyzed. Four areas of investigation have been outlined: (a) sampling error, (b) study artifacts, (c) methodological features, and (d) confounded study features.

Tests of homogeneity are often carried out in order to determine whether effect sizes are actually attributable to treatment effects or sampling error. In other words, in meta-analysis, it is necessary to find whether the various combined effect sizes all estimate the same population effect size. A homogeneous distribution suggests that a given effect size differs from the population effect size only by sampling error. In cases where a significant heterogeneity is obtained for the effect sizes, an analysis of preselected moderating variables is conducted in order to account for the observed variability in effect sizes.

Although corrections for study "artifacts" (imperfections) may be carried out, some meta-analytic

experts have argued that study imperfections should not be corrected in order to estimate what would have been found in a perfect study because these procedures are at odds with the true purpose of a meta-analysis, which is to reflect study findings accurately in a given field of research.

—Marjan Ghahramanlou-Holloway

See also Effect Size

Further Reading

- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Applied Social Research Methods Series Vol. 49). Thousand Oaks, CA: Sage.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Applied Social Research Methods Series Vol. 6). Newbury Park, CA: Sage.

The Campbell Collaboration: <http://www.campbellcollaboration.org/>

The Cochrane Collaboration: The reliable source of evidence in health care: <http://www.cochrane.org/index0.htm>

Comprehensive Meta-Analysis: A computer program for research synthesis: <http://www.meta-analysis.com/>

METRIC MULTIDIMENSIONAL SCALING

Metric multidimensional scaling (MDS) transforms a distance matrix into a set of coordinates such that the (Euclidean) distances derived from these coordinates approximate as well as possible the original distances. The basic idea of MDS is to transform the distance matrix into a cross-product matrix and then to find its eigendecomposition, which gives a principal component analysis (PCA). Like PCA, MDS can be used with supplementary or illustrative elements that are projected onto the dimensions after they have been computed.

An Example

The example is derived from O'Toole, Jiang, Abdi, and Haxby, who used a combination of principal component analysis and neural networks to analyze brain imaging data. In this study, 6 subjects were scanned using fMRI when they were watching pictures from 8 categories (faces, houses, cats, chairs, shoes, scissors, bottles, and scrambled images). The authors computed for each subject a distance matrix corresponding to how well they could predict the type of pictures that the subject was watching from his or her brain scans. The distance used was d' , which expresses the discriminability between categories.

O'Toole et al. give two distance matrices. The first one is the average distance matrix computed from the brain scans of all 6 subjects. The authors also give a distance matrix derived directly from the pictures watched by the subjects. The authors computed this distance matrix with the same algorithm that they used for the brain scans; they just substituted images for brain scans.

We will use these two matrices to review the basics of multidimensional scaling: namely, how to transform a distance matrix into a cross-product matrix and how to project a set of supplementary observations onto the space obtained by the original analysis.

Multidimensional Scaling: Eigenanalysis of a Distance Matrix

PCA is obtained by performing the eigendecomposition of a matrix. This matrix can be a correlation matrix (i.e., the variables to be analyzed are centered and normalized), a covariance matrix (i.e., the variables are centered but not normalized), or a cross-product matrix (i.e., the variables are neither centered nor normalized). A distance matrix cannot be analyzed directly using the eigendecomposition (because distance matrices are not positive semi-definite matrices), but it can be transformed into an equivalent cross-product matrix, which can then be analyzed.

Transforming a Distance Matrix Into a Cross-Product Matrix

In order to transform a distance matrix into a cross-product matrix, we start from the observation that the scalar product between two vectors can be transformed easily into a distance (the scalar product between vectors corresponds to a cross-product matrix). Let us start with some definitions. Suppose that $\mathbf{a}$ and $\mathbf{b}$ are two vectors with I elements. The Euclidean distance between these two vectors is computed as

$$d^2(\mathbf{a}, \mathbf{b}) = (\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b}). \quad (1)$$

This distance can be rewritten to isolate the scalar product between vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\begin{aligned} d^2(\mathbf{a}, \mathbf{b}) &= (\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b}) \\ &= \mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b} - 2 \times (\mathbf{a}^T\mathbf{b}), \end{aligned} \quad (2)$$

where $\mathbf{a}^T\mathbf{b}$ is the scalar product between $\mathbf{a}$ and $\mathbf{b}$.

If the data are stored into an $I \times J$ data matrix denoted $\mathbf{X}$ (where I observations are described by J variables), the between-observations cross-product matrix is then obtained as

$$\mathbf{S} = \mathbf{X} \mathbf{X}^T. \quad (3)$$

$I \times I$ $I \times J$ $J \times I$

A distance matrix can be computed directly from the cross-product matrix as

$$\mathbf{D} = \mathbf{s}\mathbf{1}^T + \mathbf{1}\mathbf{s}^T - 2\mathbf{S}. \quad (4)$$

$I \times I$ $I \times I$ $I \times I$ $I \times I$ $I \times I$

(Note that the elements of $\mathbf{D}$ gives the *squared* Euclidean distance between rows of $\mathbf{S}$.)

This equation shows that an Euclidean distance matrix can be computed from a cross-product matrix. In order to perform MDS on a set of data, the main idea is to “revert” Equation 4 in order to obtain a cross-product matrix from a distance matrix. There is one problem when implementing this idea, namely, that *different* cross-product matrices can give the *same* distance. This can happen because distances are invariant for any change of origin. Therefore, in order to revert the equation, we need to impose an origin for

the computation of the distance. An obvious choice is to choose the origin of the distance as the center of gravity of the dimensions. With this constraint, the cross-product matrix is obtained as follows.

First define a mass vector denoted $\mathbf{m}$ whose I elements give the mass of the I rows of matrix $\mathbf{D}$. These elements are all positive, and their sum is equal to 1:

$$\mathbf{m}^T \mathbf{1} = 1. \quad (5)$$

$I \times I \times 1$

When all the rows have equal importance, each element is equal to $\frac{1}{I}$.

Second, define an $I \times I$ centering matrix denoted $\mathbf{\Xi}$ (read “big Xi”) equal to

$$\mathbf{\Xi} = \mathbf{I} - \mathbf{1} \mathbf{m}^T. \quad (6)$$

$I \times I$ $I \times I$ $I \times 1$ $1 \times I$

Finally, the cross-product matrix is obtained from matrix $\mathbf{D}$ as

$$\mathbf{S} = -\frac{1}{2} \mathbf{\Xi} \mathbf{D} \mathbf{\Xi}^T. \quad (7)$$

$I \times I$

The eigendecomposition of this matrix gives

$$\mathbf{S} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \quad (8)$$

with

$$\mathbf{U}^T \mathbf{U} = \mathbf{I} \text{ and } \mathbf{\Lambda} \text{ diagonal matrix of eigenvalues.} \quad (9)$$

See the “Appendix: Proof” section toward the end of this entry for a proof.

The scores (i.e., the projection of the rows on the principal components of the analysis of $\mathbf{S}$) are obtained as

$$\mathbf{F} = \mathbf{M}^{-\frac{1}{2}} \mathbf{U} \mathbf{\Lambda}^{-\frac{1}{2}} (\text{with } \mathbf{M} = \text{diag}\{\mathbf{m}\}). \quad (10)$$

The scores have the properties that their variance is equal to the eigenvalues

$$\mathbf{F}^T \mathbf{M} \mathbf{F} = \mathbf{\Lambda}. \quad (11)$$

Example

To illustrate the transformation of the distance matrix, we will use the distance matrix derived from the brain scans given in Table 1:

$$\mathbf{D} = \begin{bmatrix} 0.00 & 3.47 & 1.79 & 3.00 & 2.67 & 2.58 & 2.22 & 3.08 \\ 3.47 & 0.00 & 3.39 & 2.18 & 2.86 & 2.69 & 2.89 & 2.62 \\ 1.79 & 3.39 & 0.00 & 2.18 & 2.34 & 2.09 & 2.31 & 2.88 \\ 3.00 & 2.18 & 2.18 & 0.00 & 1.73 & 1.55 & 1.23 & 2.07 \\ 2.67 & 2.86 & 2.34 & 1.73 & 0.00 & 1.44 & 1.29 & 2.38 \\ 2.58 & 2.69 & 2.09 & 1.55 & 1.44 & 0.00 & 1.19 & 2.15 \\ 2.22 & 2.89 & 2.31 & 1.23 & 1.29 & 1.19 & 0.00 & 2.07 \\ 3.08 & 2.62 & 2.88 & 2.07 & 2.38 & 2.15 & 2.07 & 0.00 \end{bmatrix}. \quad (12)$$

The elements of the mass vector $\mathbf{m}$ are all equal to $1/8$:

$$\mathbf{m}^T = [.125 \ .125 \ .125 \ .125 \ .125 \ .125 \ .125 \ .125]. \quad (13)$$

The centering matrix is equal to

$$\mathbf{E}_{8 \times 8} = \begin{bmatrix} .875 & -.125 & -.125 & -.125 & -.125 & -.125 & -.125 & -.125 \\ -.125 & .875 & -.125 & -.125 & -.125 & -.125 & -.125 & -.125 \\ -.125 & -.125 & .875 & -.125 & -.125 & -.125 & -.125 & -.125 \\ -.125 & -.125 & -.125 & .875 & -.125 & -.125 & -.125 & -.125 \\ -.125 & -.125 & -.125 & -.125 & .875 & -.125 & -.125 & -.125 \\ -.125 & -.125 & -.125 & -.125 & -.125 & .875 & -.125 & -.125 \\ -.125 & -.125 & -.125 & -.125 & -.125 & -.125 & .875 & -.125 \\ -.125 & -.125 & -.125 & -.125 & -.125 & -.125 & -.125 & .875 \end{bmatrix}.$$

The cross-product matrix is then equal to

$$\mathbf{S} = \begin{bmatrix} 1.34 & -0.31 & 0.34 & -0.46 & -0.25 & -0.26 & -0.12 & -0.29 \\ -0.31 & 1.51 & -0.38 & 0.03 & -0.26 & -0.24 & -0.37 & 0.02 \\ 0.34 & -0.38 & 1.12 & -0.16 & -0.19 & -0.14 & -0.27 & -0.31 \\ -0.46 & 0.03 & -0.16 & 0.74 & -0.08 & -0.05 & 0.07 & -0.09 \\ -0.25 & -0.26 & -0.19 & -0.08 & 0.83 & 0.05 & 0.09 & -0.20 \\ -0.26 & -0.24 & -0.14 & -0.05 & 0.05 & 0.71 & 0.08 & -0.15 \\ -0.12 & -0.37 & -0.27 & 0.07 & 0.09 & 0.08 & 0.65 & -0.13 \\ -0.29 & 0.02 & -0.31 & -0.09 & -0.20 & -0.15 & -0.13 & 1.15 \end{bmatrix}.$$

The eigendecomposition of $\mathbf{S}$ gives

$$\mathbf{U} = \begin{bmatrix} 0.60 & -0.36 & -0.10 & 0.48 & -0.23 & 0.02 & 0.30 \\ -0.52 & -0.64 & 0.36 & 0.14 & 0.10 & -0.06 & -0.18 \\ 0.48 & -0.17 & 0.10 & -0.67 & 0.24 & 0.04 & -0.30 \\ -0.23 & 0.16 & 0.20 & -0.38 & -0.54 & 0.29 & 0.49 \\ -0.02 & 0.39 & 0.19 & 0.28 & 0.61 & 0.47 & 0.14 \\ -0.03 & 0.32 & 0.11 & -0.00 & 0.14 & -0.83 & 0.23 \\ 0.00 & 0.38 & 0.02 & 0.25 & -0.43 & 0.04 & -0.69 \\ -0.28 & -0.08 & -0.87 & -0.09 & 0.11 & 0.04 & 0.02 \end{bmatrix} \quad (14)$$

Table 1 Matrix That Gives the d' Obtained for the Discrimination Between Categories Based On the Brain Scans

	<i>Face</i>	<i>House</i>	<i>Cat</i>	<i>Chair</i>	<i>Shoe</i>	<i>Scissors</i>	<i>Bottle</i>	<i>Scrambled</i>
Face	0.00	3.47	1.79	3.00	2.67	2.58	2.22	3.08
House	3.47	0.00	3.39	2.18	2.86	2.69	2.89	2.62
Cat	1.79	3.39	0.00	2.18	2.34	2.09	2.31	2.88
Chair	3.00	2.18	2.18	0.00	1.73	1.55	1.23	2.07
Shoes	2.67	2.86	2.34	1.73	0.00	1.44	1.29	2.38
Scissors	2.58	2.69	2.09	1.55	1.44	0.00	1.19	2.15
Bottle	2.22	2.89	2.31	1.23	1.29	1.19	0.00	2.07
Scrambled	3.08	2.62	2.88	2.07	2.38	2.15	2.07	0.00

Source: O'Toole et al. (2005). These data are obtained by averaging 12 data tables (2 per subject).

and

$$\Lambda = \begin{bmatrix} 2.22 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.72 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.23 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.00 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.79 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.69 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.39 \end{bmatrix}. \quad (15)$$

As in PCA, the eigenvalues are often transformed into percentage of explained variance (or inertia) in order to make their interpretation easier. Here, for example, we find that the first dimension “explains” 28% of the variance of the distances (i.e., $\frac{2.22}{2.22+\dots+0.39} = .28$).

We obtain the following matrix of scores:

$$\mathbf{F} = \begin{bmatrix} 2.53 & -1.35 & -0.30 & 1.36 & -0.58 & 0.04 & 0.53 \\ -2.19 & -2.37 & 1.13 & 0.39 & 0.24 & -0.15 & -0.32 \\ 2.04 & -0.63 & 0.32 & -1.90 & 0.61 & 0.10 & -0.52 \\ -0.97 & 0.61 & 0.62 & -1.09 & -1.35 & 0.68 & 0.86 \\ -0.10 & 1.44 & 0.59 & 0.81 & 1.53 & 1.10 & 0.25 \\ -0.13 & 1.18 & 0.33 & -0.00 & 0.35 & -1.96 & 0.40 \\ 0.02 & 1.41 & 0.05 & 0.70 & -1.09 & 0.09 & -1.22 \\ -1.20 & -0.29 & -2.74 & -0.27 & 0.28 & 0.10 & 0.03 \end{bmatrix}.$$

Figure 1a displays the projection of the categories on the first two dimensions. The first dimension explains 28% of the variance of the distance, it can be interpreted as the opposition of the face and cat categories to the house category (these categories are the ones most easily discriminated in the scans). The second dimension, which explains 21% of the variance, separates the small objects from the other categories.

Multidimensional Scaling: Supplementary Elements

After we have computed the MDS solution, it is possible to project supplementary or illustrative elements onto this solution. To illustrate this procedure, we will project the distance matrix obtained from the pictures (see Table 2) onto the space defined by the analysis of the brain scans.

The number of supplementary elements is denoted by I_{sup} . For each supplementary element, we need the values of its distances to all the I active elements. We

can store these distances in an $I \times I_{\text{sup}}$ supplementary distance matrix denoted $\mathbf{D}_{\text{sup}}$. So, for our example, we have

$$\mathbf{D}_{\text{sup}} = \begin{bmatrix} 0.00 & 4.52 & 4.08 & 4.08 & 4.52 & 3.97 & 3.87 & 3.73 \\ 4.52 & 0.00 & 2.85 & 4.52 & 4.52 & 4.52 & 4.08 & 4.52 \\ 4.08 & 2.85 & 0.00 & 1.61 & 2.92 & 2.81 & 1.96 & 3.17 \\ 4.08 & 4.52 & 1.61 & 0.00 & 2.82 & 2.89 & 2.91 & 3.97 \\ 4.52 & 4.52 & 2.92 & 2.82 & 0.00 & 3.55 & 3.26 & 4.52 \\ 3.97 & 4.52 & 2.81 & 2.89 & 3.55 & 0.00 & 2.09 & 3.26 \\ 3.87 & 4.08 & 1.96 & 2.91 & 3.26 & 2.09 & 0.00 & 1.50 \\ 3.73 & 4.52 & 3.17 & 3.97 & 4.52 & 3.26 & 1.50 & 0.00 \end{bmatrix}. \quad (16)$$

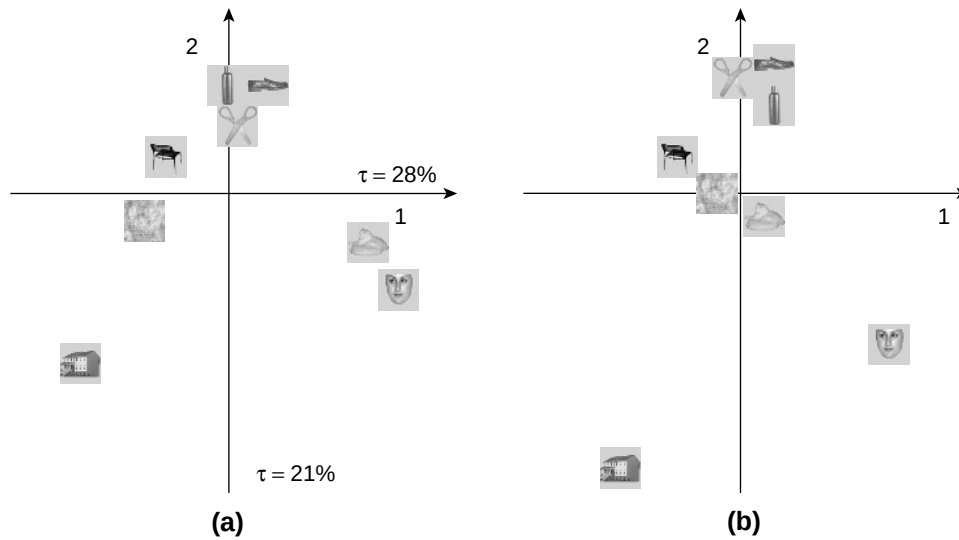


Figure 1 (a) Multidimensional Scaling of the Subjects' Distance Table; (b) Projection of the Image Distance Table as Supplementary Elements in the Subjects' Space (Distance From Tables 1 and 2)

Table 2 Matrix That Gives the d' Obtained for the Discrimination Between Categories Based On the Images Watched by the Subjects

	<i>Face</i>	<i>House</i>	<i>Cat</i>	<i>Chair</i>	<i>Shoe</i>	<i>Scissors</i>	<i>Bottle</i>	<i>Scrambled</i>
Face	0.00	4.52	4.08	4.08	4.52	3.97	3.87	3.73
House	4.52	0.00	2.85	4.52	4.52	4.52	4.08	4.52
Cat	4.08	2.85	0.00	1.61	2.92	2.81	1.96	3.17
Chair	4.08	4.52	1.61	0.00	2.82	2.89	2.91	3.97
Shoes	4.52	4.52	2.92	2.82	0.00	3.55	3.26	4.52
Scissors	3.97	4.52	2.81	2.89	3.55	0.00	2.09	3.26
Bottle	3.87	4.08	1.96	2.91	3.26	2.09	0.00	1.50
Scrambled	3.73	4.52	3.17	3.97	4.52	3.26	1.50	0.00

Source: O'Toole et al. (2005).

The first step is to transform $\mathbf{D}_{\text{sup}}$ into a cross-product matrix denoted $\mathbf{S}_{\text{sup}}$. This is done by centering the rows with the same centering matrix that was used previously to transform the distance of the active

elements. Specifically, the cross-product matrix is obtained as

$$\mathbf{S}_{\text{sup}} = -\frac{1}{2} \mathbf{E} \mathbf{D}_{\text{sup}} \mathbf{E}^T \quad (17)$$

For our example, this gives

$$\mathbf{S}_{\text{sup}} = \begin{bmatrix} 1.80 & -0.41 & -0.83 & -0.62 & -0.63 & -0.54 & -0.71 & -0.32 \\ -0.46 & 1.85 & -0.21 & -0.84 & -0.63 & -0.82 & -0.81 & -0.72 \\ -0.24 & 0.42 & 1.21 & 0.62 & 0.17 & 0.04 & 0.25 & -0.04 \\ -0.24 & -0.41 & 0.41 & 1.42 & 0.22 & 0.00 & -0.22 & -0.44 \\ -0.46 & -0.41 & -0.25 & 0.01 & 1.63 & -0.33 & -0.40 & -0.72 \\ -0.18 & -0.41 & -0.19 & -0.02 & -0.14 & 1.44 & 0.18 & -0.09 \\ -0.14 & -0.20 & 0.23 & -0.03 & 0.00 & 0.40 & 1.23 & 0.79 \\ -0.07 & -0.41 & -0.37 & -0.56 & -0.63 & -0.19 & 0.48 & 1.54 \end{bmatrix}. \quad (18)$$

The next step is to project the matrix $\mathbf{S}_{\text{sup}}$ onto the space defined by the analysis of the active distance matrix. We denote by $\mathbf{F}_{\text{sup}}$ the matrix of projection of the supplementary elements. Its computational formula is obtained by first combining Equations 10 and 8 in order to get

$$\mathbf{F} = \mathbf{S}^T \mathbf{M}^{-\frac{1}{2}} \mathbf{U} \mathbf{\Lambda}^{-\frac{1}{2}}, \quad (19)$$

and then substituting $\mathbf{S}_{\text{sup}}$ for $\mathbf{S}$ and simplifying gives

$$\mathbf{F}_{\text{sup}} = \mathbf{S}_{\text{sup}}^T \mathbf{M}^{-\frac{1}{2}} \mathbf{U} \mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{S}_{\text{sup}}^T \mathbf{F} \mathbf{\Lambda}^{-1}. \quad (20)$$

For our example, this equation gives the following values:

$$\mathbf{F}_{\text{sup}} = \begin{bmatrix} 2.45 & -1.38 & -1.18 & 2.53 & -2.05 & -0.30 & 2.56 \\ -1.46 & -3.24 & 2.30 & -0.55 & 1.04 & -0.34 & -3.68 \\ 0.43 & 0.55 & 1.19 & -3.89 & -0.23 & 0.69 & -2.81 \\ 0.37 & 1.98 & 1.50 & -3.74 & -1.90 & 1.61 & 2.21 \\ 0.25 & 2.74 & 1.87 & -0.19 & 2.90 & 3.24 & 0.79 \\ 0.26 & 2.61 & 0.06 & -1.07 & -0.44 & -4.43 & -0.11 \\ 0.08 & 2.22 & -1.79 & -1.10 & -1.37 & -0.99 & -5.05 \\ -0.30 & 0.83 & -4.59 & -0.59 & -1.24 & -0.87 & -3.72 \end{bmatrix}. \quad (21)$$

Figure 1b displays the projection of the supplementary categories on the first two dimensions. Comparing plots *a* and *b* shows that an analysis of the pictures reveals a general map very similar to the analysis of the brain scans with only one major difference: The cat category for the images moves to the center of the space. This suggests that the cat category

is interpreted by the subjects as being facelike (i.e., “cats have faces”).

Analyzing Nonmetric Data

Metric MDS is adequate only when dealing with distances. In order to accommodate weaker measurements

(called dissimilarities), nonmetric MDS is adequate. It derives a Euclidean distance approximation using only the ordinal information from the data.

Appendix: Proof

We start with an $I \times I$ distance matrix $\mathbf{D}$ and an $I \times 1$ vector of mass (whose elements are all positive or zero and whose sum is equal to 1) denoted $\mathbf{m}$ and such that

$$\mathbf{m}^T \mathbf{1}_{I \times 1} = 1. \quad (22)$$

The centering matrix is equal to

$$\mathbf{E} = \mathbf{I}_{I \times I} - \mathbf{1}_{I \times 1} \mathbf{m}^T_{1 \times I}. \quad (23)$$

We want to show that the following cross-product matrix,

$$\mathbf{S}_{I \times I} = -\frac{1}{2} \mathbf{E} \mathbf{D} \mathbf{E}^T, \quad (24)$$

will give back the original distance matrix when the distance matrix is computed as

$$\mathbf{D}_{I \times I} = \mathbf{s} \mathbf{1}^T_{I \times 1} + \mathbf{1} \mathbf{s}^T_{1 \times I} - 2\mathbf{S}_{I \times I}. \quad (25)$$

In order to do so, we need to choose an origin for the coordinates (because several coordinates systems will give the same distance matrix). A natural choice is to assume that the data are centered (i.e., the mean of each original variable is equal to zero). There, we assume that the mean vector, denoted $\mathbf{c}$, is computed as

$$\mathbf{c}_{J \times 1} = \mathbf{X}^T_{J \times I} \mathbf{m}_{I \times 1} \quad (26)$$

(for some data matrix $\mathbf{X}$). Because the origin of the space is located at the center of gravity, its coordinates are equal to $\mathbf{c} = \mathbf{0}$. The cross-product matrix can therefore be computed as

$$\begin{aligned} \mathbf{S}_{I \times I} &= \left(\mathbf{X}_{I \times J} - \mathbf{1}_{I \times 1} \mathbf{c}^T_{1 \times J} \right) \left(\mathbf{X}_{I \times J} - \mathbf{1}_{I \times 1} \mathbf{c}^T_{1 \times J} \right)^T \\ &= \left(\mathbf{X}_{I \times J} - \mathbf{1}_{I \times 1} \mathbf{c}^T_{1 \times J} \right) \left(\mathbf{X}^T_{J \times I} - \mathbf{c} \mathbf{1}^T_{1 \times I} \right) \end{aligned} \quad (27)$$

First, we assume that there exists a matrix denoted $\mathbf{S}$ such that Equation 25 is satisfied. Then we plug Equation 25 into Equation 24, develop, and simplify in order to get

$$-\frac{1}{2} \mathbf{E} \mathbf{D} \mathbf{E}^T = -\frac{1}{2} \mathbf{E} \mathbf{s} \mathbf{1}^T \mathbf{E}^T - \frac{1}{2} \mathbf{E} \mathbf{1} \mathbf{s}^T \mathbf{E}^T + \mathbf{E} \mathbf{S} \mathbf{E}^T. \quad (28)$$

Then, we show that the terms $\mathbf{E}(\mathbf{s} \mathbf{1}^T) \mathbf{E}^T$ and $\mathbf{E}(\mathbf{1} \mathbf{s}^T) \mathbf{E}^T$ are null because

$$\begin{aligned} (\mathbf{s} \mathbf{1}^T) \mathbf{E}^T &= \mathbf{s} \mathbf{1}^T (\mathbf{I} - \mathbf{1} \mathbf{m}^T)^T \\ &= \mathbf{s} \mathbf{1}^T (\mathbf{I} - \mathbf{m} \mathbf{1}^T) \\ &= \mathbf{s} \mathbf{1}^T - \mathbf{s} \mathbf{1}^T \mathbf{m} \mathbf{1}^T \text{ (but from Equation 22: } \mathbf{1}^T \mathbf{m} = 1) \\ &= \mathbf{s} \mathbf{1}^T - \mathbf{s} \mathbf{1}^T \\ &= \mathbf{0}_{I \times 1}. \end{aligned} \quad (29)$$

The last thing to show now is that the term $\mathbf{E} \mathbf{S} \mathbf{E}^T$ is equal to $\mathbf{S}$. This is shown by developing

$$\begin{aligned} \mathbf{E} \mathbf{S} \mathbf{E}^T &= (\mathbf{I} - \mathbf{1} \mathbf{m}^T) \mathbf{S} (\mathbf{I} - \mathbf{m} \mathbf{1}^T) \\ &= \mathbf{S} - \mathbf{S} \mathbf{m} \mathbf{1}^T - \mathbf{1} \mathbf{m}^T \mathbf{S} + \mathbf{1} \mathbf{m}^T \mathbf{S} \mathbf{m} \mathbf{1}^T. \end{aligned} \quad (30)$$

Because

$$\begin{aligned} (\mathbf{X}^T - \mathbf{c} \mathbf{1}^T) \mathbf{m} &= \mathbf{X}^T \mathbf{m} - \mathbf{c} \mathbf{1}^T \mathbf{m} \text{ (cf. Equations 26 and 22)} \\ &= \mathbf{c} - \mathbf{c} \\ &= \mathbf{0}_{1 \times 1}, \end{aligned} \quad (31)$$

we get (cf. Equation 27):

$$\mathbf{S} \mathbf{m} = (\mathbf{X} - \mathbf{1} \mathbf{c}^T) (\mathbf{X}^T - \mathbf{c} \mathbf{1}^T) \mathbf{m} = \mathbf{0}. \quad (32)$$

Therefore, Equation 33 becomes

$$\mathbf{E} \mathbf{S} \mathbf{E}^T = \mathbf{S}, \quad (33)$$

which leads to

$$-\frac{1}{2} \mathbf{E} \mathbf{D} \mathbf{E}^T = \mathbf{S}, \quad (34)$$

which completes the proof.

—Hervé Abdi

See also Centroid; Distance; Eigendecomposition; Principal Component Analysis; Signal Detection Theory; Singular and Generalized Singular Value Decomposition; STATIS

Further Reading

- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer Verlag.
- O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in the ventral temporal cortex. *Journal of Cognitive Neuroscience*, 17, 580–590.
- Shepard, R. N. (1966). Metric structure in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.
- Togerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

MILLON BEHAVIORAL MEDICINE DIAGNOSTIC

The Millon Behavioral Medicine Diagnostic (MBMD), published by NCS Assessments in 2001, is a revision of the 1974 Millon Behavioral Health Inventory (MBHI). Both are designed to evaluate psychosocial factors that influence the course of medical disease and to assist health care providers in planning successful treatment regimens. The MBMD eliminated underperforming scales of the MBHI, widened the base of medical disorders, and added new scales to assist case management, including scales responsive to *DSM-IV* revisions concerning psychological factors affecting medical conditions. Recent Stress and Chronic Tension subscales were eliminated from the revision even though both factors play major roles in primary, secondary, and tertiary prevention.

The MBMD contains 165 true/false self-report items that contribute to 39 scales that inform about Response Patterns (Disclosure, Desirability, Debasement); Negative Health Habits (Alcohol, Drugs, Eating, Caffeine, Inactivity, Smoking); Psychiatric Indicators (Anxiety-Tension, Depression, Cognitive Dysfunction, Emotional Lability, Guardedness); Coping Styles (Introverted, Inhibited, Dejected, Cooperative, Sociable, Confident, Nonconforming, Forceful, Respectful, Oppositional, Denigrating); Stress Moderators (Spiritual Absence, Illness Apprehension, Functional Deficits, Pain Sensitivity, Social Isolation, Future Pessimism); Prognostic Indicators (Interventional Fragility, Medication

Abuse, Information Discomfort, Utilization Excess, Problematic Compliance); Management Guides (Adjustment Difficulties, Psychiatric-Psychosocial Referral); and a Validity Indicator.

The MBMD is administered within 20 to 25 minutes for medical patients between 18 and 85 years old. Interpretation of the MBMD is based on prevalence scores rather than traditional scaling methods. This approach is intended to increase certainty about patient evaluation. Prevalence score norms are generic, although there is no reason to believe that prevalence scores are the same across medical disorders. Test-retest reliability was assessed at 7- and 30-day intervals and resulted in moderate to strong coefficients (.71 to .92). Most scales obtained a sound Cronbach's alpha coefficient (.74 to .89). However, on approximately half of the Coping Styles scales, lower coefficients were obtained (.54 to .67). Redundancy of items across 39 subscales is problematic; that is, 165 items are scored 492 times.

The MBMD manual presents concurrent validity through comparison of the individual scales against other measures with well-established validity, such as the State-Trait Anxiety Inventory, Profile of Mood States, and Beck Depression Inventory. Moderate to high (.50 to .87) convergent correlations between the MBMD and other standard measures were found, excluding the Problematic Compliance (.38) and Utilization Excess (.39 to .52) scales. Predictive validity studies have shown scores on various scales to be valid for predicting disease adjustment and adaptations, probability of seeking treatment, and the course of medical illness. The extreme revision of the MBMD questions the applicability of research that accrued with the MBHI. Rather, a new empirical base, presently small but growing, will help guide practitioners in the application of test results.

The MBMD is a novel tool for assessing the complexity of identifying the treatment needs of medical patients. This tool lends hope for the further integration of the biopsychosocial model into traditional medical systems, providing a better understanding of the prognostics, treatment, rehabilitation, and management of disease.

—Joseph A. Doster and Lyndsi M. Grover

Further Reading

- Bockian, N., Meager, S., & Millon, T. (2000). Assessing personality with the Millon Behavioral Health Inventory, the Millon Behavioral Medicine Diagnostic, and the Millon Clinical Multiaxial Inventory. In R. J. Gatchel & J. N. Weisberg (Eds.), *Personality characteristics of patients with pain* (pp. 61–88). Washington, DC: American Psychological Association.
- Millon, T. (1990). *Toward a new personology*. New York: Wiley.
- Millon, T., Green, C. J., & Meagher, R. B. (1979). The MBHI: A new inventory for the psychodiagnostician in medical settings. *Professional Psychology: Research and Practice*, 10, 529–539.

MILLON CLINICAL MULTIAXIAL INVENTORY-III

The Millon Clinical Multiaxial Inventory-III (MCMI-III) is a 170-item objective measure of personality developed by Theodore Millon. It is somewhat unique among self-report inventories in that it focuses on personality disorders and their symptoms rather than standard clinical disorders such as depression and anxiety, and/or normal variations in personality. In fact, the focus is exclusively on clinical populations, and it should not be used with normal populations. First published in 1977, the MCMI has undergone two revisions (currently MCMI-III) and currently reflects a combination of Millon's own theory of personality disorders and the *DSM-IV* criteria for diagnosis of those disorders. It is ranked second only to the Minnesota Multiphasic Personality Inventory (MMPI-2) in frequency of use in clinical settings for personality assessment.

There are essentially three broad categories of scores generated on the MCMI: 4 Validity scales, 14 Personality Disorder scales (3 of which are labeled Severe), and 10 Clinical Syndrome scales (3 of which are labeled Severe). The Validity scales can help to determine the degree to which an individual is minimizing or exaggerating symptoms. The Personality scales cover the 10 personality disorders listed in the *DSM-IV* and four others, two of which are listed in the *DSM-IV* for further study and two that are derived

from Millon's theory of personality. The Clinical Syndrome scales are composed of three that assess disorders of mood, two that focus on anxiety, two that deal with disorders of belief and thought, one each for drug and alcohol problems, and one somatoform scale. All of these scales use a base rate scoring system that adjusts scores based on how frequently the symptoms measured will occur in a clinical population. In this way, disorders whose symptoms are somewhat common in a clinical population will require a higher level of symptom reporting than do disorders whose symptoms are more rare in order to be considered elevated. Two cut-score points are established (75BR suggesting the presence of symptoms, and 85BR suggesting the presence of the disorder). Regardless of the level of elevation, it is important that the clinician using the test incorporates additional information into any diagnostic decision, and should never base a diagnosis on the test results in isolation.

Millon has also developed inventories for use with other populations. The Millon Behavioral Medicine Diagnostic (MBMD) is an updated revision of the earlier Millon Behavioral Health Inventory (MBHI) and is designed for use with adults in primary care or other medical settings. The Millon Adolescent Clinical Inventory (MACI) is essentially a version of the MCMI for use with 13- to 19-year-olds, whereas the Millon Pre-Adolescent Clinical Inventory (M-PACI) extends coverage to 9- to 12-year-olds. The Millon Adolescent Personality Inventory (MAPI) combines assessment of normal personality factors with the MACI, and finally, the Millon Inventory of Personality Styles-Revised (MIPS-Revised) is an assessment of normal personality traits in adults. Hand scoring, mail-in scoring, and on-site scoring software is available for all forms, as are both scoring and interpretive reports. Clinicians are charged a per-use fee, with the exception of hand scoring. However, because hand scoring is complicated, time-consuming, and error prone, it is not recommended.

—Steve Saladin

See also Basic Personality Inventory; Comrey Personality Scales; Minnesota Multiphasic Personality Inventory

Further Reading

Jankowski, D., & Millon, T. (2002). *A beginner's guide to the MCMI-III*. Washington, DC: American Psychological Association.

Strack, S. (1999). *Essentials of Millon inventories assessment*. New York: Wiley.

Millon Clinical Multiaxial Inventory-III: <http://hometown.net/MCMI.htm>

MINNESOTA CLERICAL TEST

The Minnesota Clerical Test (published by The Psychological Corporation) measures perceptual abilities that are commonly used in the performance of various clerical activities. Most often, the MCT is used by employers to select employees who can work with information quickly and accurately, for clerical positions such as clerks, cashiers, tellers, and typists.

Composed of two parts—Number Comparison and Name Comparison—the test assesses the accuracy and speed with which test takers are able to distinguish between pairs of numbers or names that are similar and those that are identical. For example, an item including two numbers that are similar (e.g., 12343 and 13243) would be correctly identified as being different, whereas two numbers that are identical (e.g., 12232 and 12232) would be correctly identified as being the same. An item including two similar names (e.g., Davenport Inc. and Davinport Inc.) would be correctly identified as different, whereas two identical names (e.g., Smith & Co. and Smith & Co.) would be correctly identified as the same.

After receiving the test materials, as well as completing identification information and sample items, the test taker is given 15 minutes to complete both parts. Each consisting of 200 items, the two parts are scored by subtracting the number of incorrect responses from the number of correct responses.

The MCT was constructed in 1931 by psychologists at the University of Minnesota and first published in 1933. Initially known as the Minnesota Vocational Test for Clerical Workers, the test was given its current name in 1946. Remarkably, the

content of the test's items have not been altered since the original publication, although the manual has been revised five times with particular regard to psychometric and normative information. The most recent manual presents normative data, which enables a comparison of raw scores with larger groups, for trainees and employees working at utility companies, banks, financial institutions, universities, and temp agencies, as well as male, female, Caucasian, African American, and Hispanic individuals.

Overall, the MCT has been found to have good psychometric properties. In particular, test scores predicted future levels of job performance in clerical occupations and were related to typing speed, General Clerical Test scores, and Wechsler Adult Intelligence Scale scores. Also, the relationships found between the number and name comparison parts suggest that the two parts are measuring similar but distinct abilities. Although the MCT is used primarily in vocational settings, the test also is used by counselors to aid those considering clerical occupations, and by researchers in areas such as cognitive and vocational psychology.

—Shawn T. Bubany and Jo-Ida C. Hansen

Further Reading

Andrew, D. M., Paterson, D. G., & Longstaff, H. P. (1979). *Manual for the Minnesota Clerical Test*. New York: Psychological Corporation.

Hall, W. B., & Gough, H. G. (1977). Selecting statistical clerks with the Minnesota Clerical Test. *Journal of Psychology*, 96, 297–301.

Super, D. E., & Crites, J. O. (1962). *Appraising vocational fitness*. New York: Harper & Brothers.

Clerical-type work: www.bls.gov/oco/ocos130.htm#nature

The Psychological Corporation: <http://www.psychopr.com>

MINNESOTA MULTIPHASIC PERSONALITY INVENTORY

The Minnesota Multiphasic Personality Inventory (MMPI) is the most widely used psychological test of

all time. It is routinely administered to patients in psychiatric hospitals and mental health clinics, students in college counseling centers, individuals involved in the criminal justice system, and applicants for certain areas of employment (and some graduate school programs). It is suitable for use with adults aged 18 and older, and there is an adolescent version (MMPI-A) for ages 14 to 18.

The MMPI was first published in 1943 by Starke Hathaway and J. C. McKinley of the University of Minnesota and revised in 1989 to become the MMPI-2. It is currently published by the University of Minnesota Press and marketed through Pearson Assessments. The MMPI-2 consists of 567 declarative statements that the examinee indicates as either true or false about herself or himself and takes approximately 60 to 90 minutes to complete. The test was originally designed as an objective measure for obtaining clinical-diagnostic information in psychiatric and general medical settings. Although the original clinical scales no longer match our current understanding of psychopathology, the vast amount of research on the MMPI and the MMPI-2 can lead to an enhanced understanding of how the individual is functioning (i.e., people with particular patterns to their scores tend to display particular characteristics). A skilled clinician can garner information about symptoms, personality, relationship style, response to stress, level of distress, self-perception, and many other aspects of functioning, and then can facilitate treatment planning and/or differential diagnosis.

The MMPI-2 generates a variety of scores, including the 10 original clinical scales, 9 restructured clinical scales, 8 validity scales, 15 content scales, subscales for many of these scales, and a host of supplementary scales. These scales cover areas ranging from depression and anxiety to self-esteem and openness to treatment to gender identity and level of distress. The sheer volume of information is often confusing and overwhelming to the novice. In response to this, there are a number of computerized interpretive programs on the market, some targeted toward a particular setting (e.g., forensic, personnel). The use of these reports has dramatically increased in recent years, but remains somewhat controversial among the psychological community because they cannot take into consideration the

individual's history and current circumstances. Such reports typically have a warning stating that the content should serve as hypotheses that a trained clinician would use as a starting point.

—Steve Saladin

See also Basic Personality Inventory; Comrey Personality Scales; NEO Personality Inventory

Further Reading

- Buthcher, J. N. (2004). *A beginner's guide to the MMPI-2*. Washington, DC: American Psychological Association.
- Friedman, A. J., Lewak, R., Nichols, D. S., & Webb, J. T. (2001). *Psychological assessment with the MMPI-2*. Mahwah, NJ: Erlbaum.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Minnesota Multiphasic Personality Inventory article: <http://en.wikipedia.org/wiki/MMPI>

MMPI-2 information site: <http://www.mmapi-info.com/mmapi tart.html>

MISSING DATA METHOD

In 1959, Robert E. Dear proposed a missing data method that used the first extracted principle component. This unique method appeared in a technical report produced by the now-defunct Systems Development Corporation (SDC) in Santa Monica, California. For the next two decades, the Dear missing data method was a subject of much discussion and research. This report is no longer in print and its availability is scarce. It is, however, an important method that was documented in Neil Timm's doctoral dissertation and subsequent *Psychometrika* article in 1970. Terry Gleason and Richard Staelin also discussed it in their 1975 *Psychometrika* article. These publications document the efficiency and statistical properties of the Dear method. They also compare it to other available methods such as the ones proposed by Buck and Wilks. Both Timm and Gleason and Staelin found the Dear method to be among the best methods in terms of computational ease and accuracy.

Dr. Dear created this method because of the problems that researchers at SDC encountered using multiple regression with missing data values. A typical illustration of the missing data problem involves the prediction of a certain health variable for a sample of patients given a set of fixed, independent variables that should have been recorded for each patient in the study. The set of independent variables might include case history items. These would include the patient's age when she or he may have acquired a given disease, the age at death of each parent, the number of siblings who may have acquired a given disease that has a hereditary basis, and so forth. The set of independent variables may also include clinical measurements such as the patient's blood type, height, and rate of basal metabolism. In carrying out analyses of the dependence of some health variable on a set of case history and clinical measurement variables, it is not uncommon to find that not all of the variables are recorded for each of the patients. For some of the patients in the study, some of the case history items may not have been obtainable or were not recorded, and some of the clinical measurements may not have been made or were not recorded. In such cases, if these patients' data are to be included in the study, some method must be used for estimating the missing items.

Dr. Dear studied psychometrics and statistics with Professor Paul Horst at the University of Washington. One of the foundational cornerstones of the Dear method is from Horst's earlier work on estimation when data values were missing. Both Timm and Gleason and Staelin have documented the performance of the Dear method in estimating missing data values. These studies compared it to other methods of handling missing data in terms of efficiency and bias. It has fared well enough that researchers who want to estimate missing data values should consider it. The method is not difficult from a conceptual and computational point of view.

Although Dear's paper was concerned with the "estimation" of values to use for missing independent variables in multiple regression models, the method could be applied to nonregression types of problems where individual data points need to be estimated.

Dear's method, however, handles missing data values for independent variables that are considered as fixed numbers or nonrandom variables.

In multiple regression, the real statistical estimation problem is the estimation of regression coefficients for the regression of a dependent variable on a set of independent variables. However, Dear points out that there are occasions when missing values among the independent variables necessitate methods for constructing or estimating entries for these missing data before the main regression analysis can be carried out.

The Regression Model

The multiple regression model is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_m X_{mi} + \varepsilon_i$$

or

$$Y_i = \sum_j^n \beta_j X_{ij} + \varepsilon_i, \text{ where } i = 1, 2, \dots, m.$$

In this model, the Y s are the observable dependent variable values (Y is assumed to be the random variable). The X s are the fixed, observable independent variables. The coefficients, β s, are the regression parameters to be estimated, and the random variables, ε s, are the errors or differences between observed and predicted Y values.

The only assumptions that are made about the random variables of this model are (a) the expected value of the errors is zero, (b) the expected value of covariance between error terms is zero, and (c) the expected value of each error value squared is some fixed value σ^2 across all values of i .

If $X_{ij}^{(k)}$ represents the values where X is known and $X_{ij}^{(u)}$ represents the values where X is unknown, the regression model equation can be written as

$$Y_i = \sum_j^n \beta_j X_{ij}^{(k)} + \sum_j^n \beta_j X_{ij}^{(u)} + \varepsilon_i.$$

Using matrices, the equation becomes

$$\mathbf{Y} = \mathbf{X}^{(k)}\mathbf{B} + \mathbf{X}^{(u)}\mathbf{B} + \mathbf{E}.$$

The description of the missing-data model explicitly requires the rewriting of the $\mathbf{X}$ matrix in an expanded form and the definition of several auxiliary weighting matrices.

The regression model can now be written in terms of the newly defined matrices as

$$\mathbf{Y} = (\mathbf{I}' - \mathbf{C})' \mathbf{X}^* \mathbf{B} + \mathbf{C}' \mathbf{X}^* \mathbf{B} + \mathbf{E},$$

where $\mathbf{I}$ is the identity matrix with dimensions $(mn \times m)$ and $\mathbf{C}$ is the matrix of indicator variables with dimensions $(mn \times m)$. The missing values will be accounted for by elements of the $\mathbf{C}$ matrix. The indicator variables, c_{ij} , are specified, where

c_{ij} is 0 if the value for X_{ij} is known,

c_{ij} is 1 if the value for X_{ij} is unknown (missing).

We can rewrite some terms that will assist in setting up the missing data method:

$$\mathbf{X}^{(k)} = (\mathbf{I}^* - \mathbf{C})' \mathbf{X}^* \text{ and } \mathbf{X}^{(u)} = \mathbf{C}' \mathbf{X}^*.$$

A Principal-Component Missing-Data Method

This method involves developing a certain kind of average for each person or element in the total sample of observations using the available independent variables. A prediction coefficient will also be developed for each of the independent variables to reflect differences among these variables in scale values. The missing elements in the independent variable matrix will be represented as products of these average scores for persons and the prediction coefficients for the available independent variables.

Let the mathematical model for this missing-data method be

$$\mathbf{X}^{(k)} = \mathbf{A} \mathbf{P}' + \mathbf{G}.$$

$\mathbf{X}^{(k)}$ is an $(m \times n)$ matrix. The $\mathbf{A}$ matrix is $(m \times 1)$, the $\mathbf{P}$ matrix is $(n \times 1)$, and the $\mathbf{G}$ matrix is $(m \times n)$.

The $\mathbf{A}$ matrix consists of a special kind of average constructed from the available independent variables

for each sample element. The $\mathbf{P}$ matrix consists of the prediction coefficients for each of the independent variables, and the $\mathbf{G}$ matrix consists of errors or discrepancies between the matrix of observed values of independent variables, $\mathbf{X}^{(k)}$, and the matrix of predicted values, $\mathbf{A} \mathbf{P}'$.

To find the roots and the subsequent estimates for the missing values, we first form the $\mathbf{X}^{(k)'} \mathbf{X}^{(k)}$ symmetric matrix. We then find the eigenvalues (characteristic roots) for this matrix. A number of different versions of computer programs are available to solve for this characteristic root. In the example given in this write-up, the MATRIX function within SPSS is used. The vector associated with the largest characteristic root is chosen for the vector labeled $\tilde{\mathbf{P}}$. The values of $\tilde{\mathbf{P}}$ are used to obtain the other set of least-squares estimates, $\tilde{\mathbf{A}}$.

Least-squares matrices of estimates, $\tilde{\mathbf{A}}$, are determined using the following formula:

$$\tilde{\mathbf{A}} = \frac{1}{\tilde{\mathbf{P}}' \tilde{\mathbf{P}}} \mathbf{X}^{(k)} \tilde{\mathbf{P}}.$$

The values within vectors $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}$ are then used to set up two matrices, $\mathbf{A}_u$ and $\mathbf{P}_u$, whose structures are

$$\mathbf{A}_u = \begin{bmatrix} \tilde{a}_1 & 0 & \cdots & 0 \\ \tilde{a}_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \tilde{a}_m & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \tilde{a}_1 \\ 0 & 0 & \cdots & \tilde{a}_2 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \tilde{a}_m \end{bmatrix}$$

$$\mathbf{P}_u = \begin{bmatrix} \tilde{p}_1 & 0 & \cdots & 0 \\ 0 & \tilde{p}_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \tilde{p}_n \end{bmatrix}.$$

[illegible]

$$\mathbf{A}_u = \begin{bmatrix} 3.946 & 0 & \dots & 0 \\ 3.501 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 4.508 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 3.946 \\ 0 & 0 & \dots & 3.501 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 4.508 \end{bmatrix}$$

$$\mathbf{P}_u = \begin{bmatrix} .5884 & 0 & 0 \\ 0 & .6921 & 0 \\ 0 & 0 & .4179 \end{bmatrix}$$

—Howard B. Lee

See also Data Collection; Data Mining; Secondary Data Analysis

Further Reading

- Dear, R. E. (1959). *A principal-components missing data method for multiple regression models* (Technical Report SP-86). Santa Monica, CA: Systems Development Corporation.
- Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, 40(2), 229–252.
- Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*, 35(4), 417–437.

MIXED MODELS

A mixed-effects model, or just mixed model, is a statistical model in which the set of predictor variables includes both fixed and random effects. A common application of mixed models is to longitudinal, or repeated measures, data in which the data consist of multiple observations on each subject. Here, observations on different subjects can be treated as independent, but observations on the same subject cannot. The inclusion of random effects in the model induces

a correlation among repeated measures on a given subject. More generally, mixed-effects models are potentially useful any time a practitioner is faced with grouped data, in which observations are correlated within groups but independent across groups.

Whether a predictor term is to be treated as a fixed or random effect depends on the desired scope of inference, as well as the mechanism by which factor levels were chosen for inclusion in the study. If the investigator is concerned only with the factor levels included in the data set, then that factor should be treated as a fixed effect. If the investigator wishes to draw inference about the population from which the observed levels were drawn, the factor should be modeled as a random effect. Alternatively, the analyst can ask, If the experiment were to be repeated, would the observed levels of this factor be the same or possibly different? If the same, then the factor is a fixed effect; otherwise, it should be treated as a random effect.

The mixed model most commonly encountered in applications is the linear mixed-effects (LME) model with normally distributed random effects. A mathematical formulation of the model is given here. Let y_i denote the m_i -dimensional vector of responses for the i th subject in the study, for i going from 1 to M , the total number of subjects. Under the LME model, we have

$$y_i = X_i\beta + Z_ib_i + e_i,$$

where X_i and Z_i are known matrices of dimension m_i by p and m_i by q respectively, β is the p -dimensional vector of unknown fixed effects, b_i is the q -dimensional vector of random effects associated with the i th subject, and e_i is a random error term. The random vectors b_i and e_i , for i going from 1 to M , are assumed to be mutually independent and multivariate normally distributed with a mean of zero. The unknown parameters in this model consist of the vector of coefficients β and whatever parameters determine the covariance matrices of the random vectors b_i and e_i . The usual statistical inference consists of the estimation of unknown parameters and prediction of the unobserved random effects.

Following are two examples of the general LME model. Both data sets are included in the “nlme” library within the free statistical software package R.

Example 1: An Ergometric Experiment

This experiment involved nine human subjects and four types of stools. Each subject was asked to arise from each of the stools, and the effort exerted (Borg scale) was recorded. The data are available with the R package, as the object “ergoStool” in the “nlme” library.

Let y_{ij} denote the effort exerted by the i th subject in arising from the j th stool—that will be the response variable. The model should include two categorical predictors: subject (9 levels) and stool type (4 levels). Are they fixed or random effects? The four stool types were selected by the experimenters; thus, stool type is a fixed effect. As for the subjects, there is no particular interest in these nine individuals; rather, the goal is to study the person-to-person variability in effort required by all persons. The nine subjects are just a random sample from some human population, so subject is a random effect. Another way to reach the same conclusion is to note that if the experiment were repeated, the same four stool types would be used because they are part of the experimental design. But another random sample would yield a different set of nine individuals. Thus, stool type is a fixed effect and subject is a random effect.

A reasonable LME model for this experiment is given by

$$y_{ij} = \beta_j + b_i + e_{ij},$$

where the subject random effects $b_1, \dots, b_9$ are independent and identically distributed (i.i.d.) as normal with mean 0 and variance σ_b^2 , and the random error terms e_{ij} are i.i.d. as normal with mean 0 and variance σ_e^2 . The model parameters are the four stool means $\beta_1, \beta_2, \beta_3$, and β_4 , and the variance components σ_b^2 and σ_e^2 .

One of the reasons mixed-effects models are so widely applicable is that they naturally and elegantly induce correlation among repeated measures on the same subject. For the model given here, it can be

shown that this correlation is given by $\sigma_b^2/(\sigma_b^2 + \sigma_e^2)$, which, depending on the values of σ_b^2 and σ_e^2 , can be anything between 0 and 1. Of course, the correlation between measurements on different subjects is 0, as the subjects are assumed to be independent.

In Example 1, we assume an additive effect (i.e., no interaction) between subject and stool type. This may be an unreasonable assumption, but it is necessary for these data because there is only one observation at each subject-stool type combination. The experiment in Example 1 is said to be unreplicated. For replicated experiments, an interaction term can be included in the model. Of course, the interaction between a fixed effect and a random effect is a random effect, as the following example illustrates.

Example 2: An Industrial Experiment

Six workers in a plant were randomly selected, and each worker was tested on each of three different machine types. The data are summarized in an interaction plot (see Figure 1). Each point in the graph represents the average score for a particular worker-machine combination—18 such points total. The position on the horizontal axis indicates the machine, and the six lines correspond to the six workers. That the

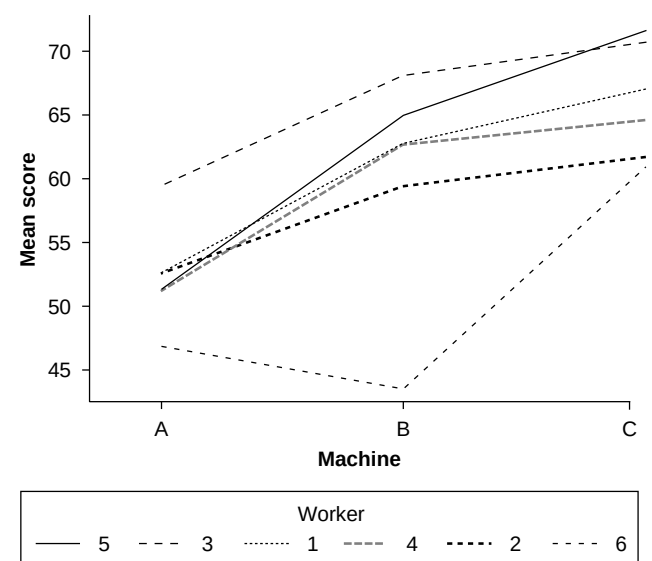


Figure 1 Interaction Plot for Example 2, Suggests Machine by Worker Interaction Effect

lines are not parallel indicates the presence of an interaction effect in predicting the response score from the predictor variables machine and worker.

Employing the same logic as in Example 1, we conclude that machine is a fixed effect and worker a random effect—the three machine types were precisely those the experimenter wished to compare, whereas the six workers were drawn at random from a population of plant workers. What about the interaction term? Is that a fixed or a random effect? There are 18 levels of worker-machine combination, and they are only a random sample of those that might have been observed. A different random sample of workers would have yielded different observed levels of worker-machine combination, and thus the interaction term is random. This result is true in general: The interaction between a fixed main effect and a random main effect is a random effect.

Let y_{ijk} denote the score for worker i on machine j in trial k , where i runs from 1 to 6 and j and k go from 1 to 3. A common model for this experiment is given by

$$y_{ijk} = \beta_j + b_i + c_{ij} + e_{ijk},$$

where the random worker effects b_i , the random interaction effects c_{ij} , and the random error terms e_{ijk} are mutually independent, normally distributed with mean 0 and variances σ_b^2 , σ_c^2 , and σ_e^2 , respectively.

According to this model, the scores of a particular worker are positively correlated, with a higher correlation between two tests on the same machine than between tests on different machines. More precisely, it can be shown that the correlation between two tests of the same worker on the same machine is equal to $(\sigma_b^2 + \sigma_c^2)/(\sigma_b^2 + \sigma_c^2 + \sigma_e^2)$, and the correlation between two tests of the same worker but on different machines is equal to $\sigma_c^2/(\sigma_b^2 + \sigma_c^2 + \sigma_e^2)$. Of course, the scores of two different workers are uncorrelated because the workers were tested independently in the experiment.

Estimation

In fitting mixed-effects models to data, we distinguish between the estimation of unknown parameters and

the prediction of unobserved random effects. The parameters in an LME model include the fixed-effects coefficients (the β_j in the examples) and the variance-covariance parameters (the σ^2 terms in the examples). These are unknown but fixed quantities, and they can be estimated from the data using one of the methods discussed below. The unobserved random effects are likewise unknown, and the investigators may wish to make an intelligent guess of their value. Because these are random quantities, this is a problem not of estimation but of prediction.

The two most common approaches to estimation are the methods of maximum likelihood (ML) and restricted maximum likelihood (REML). Both methods are available in the R software, in function “lme” (library “nlme”), with REML as the default method. Some authors have noted the tendency for ML to underestimate the variance components, and thus we will focus on REML in the example.

Example 1 Continued: Estimation

The R commands needed to produce REML estimates are given here. First, we must create an “lme object”—we call it “ergoStool.lme.” On the R command line, enter “library(nlme)” to load the “nlme” library, then “ergoStool.lme <- lme (effort~Type-1, data=ergoStool, random=~1|Subject)” to create the object. Entering the name of the object on the command line produces the following output:

```
> ergoStool.lme
Linear mixed-effects model fit by REML
Data: NULL
Log-restricted-likelihood: -60.5654
Fixed: effort ~ Type - 1
TypeT1 TypeT2 TypeT3 TypeT4
8.555556 12.444444 10.777778 9.222222

Random effects:
Formula: ~1 | Subject
(Intercept) Residual
StdDev: 1.332465 1.100295
```

```
Number of Observations: 36
Number of Groups: 9
```

Thus, the REML estimates of β_1 , β_2 , β_3 , and β_4 are 8.56, 12.44, 10.78, and 9.22, respectively. The standard deviations σ_b and σ_e are estimated as 1.33 and 1.10, respectively.

Prediction

The usual method of predicting unobserved random effects is based on an empirical Bayes approach. The method produces the Best Linear Unbiased Predictors, or BLUPs. In R, BLUPs can be obtained from the function “lme,” as we now demonstrate.

Example 1 Continued: Prediction

Once an “lme object” is created in R, the command “random.effects” produces BLUPs of the random effects. For the stool data, the BLUPs based on REML estimates are as follows:

```
> random.effects(ergoStool.lme)
(Intercept)
8 -1.708716e+00
5 -1.495127e+00
4 -8.543581e-01
9 -2.135895e-01
6 -1.912893e-15
3 4.271791e-01
7 4.271791e-01
1 1.708716e+00
2 1.708716e+00
```

We see that Subject 8 rose from the stools with the least effort, and Subject 2 required the greatest effort. BLUPs based on the ML estimates are similar.

Other Mixed-Effects Models

We have thus far discussed the inclusion of random effects only in linear statistical models. Generalized linear models (GLMs), such as the logistic regression model, can also have random effects among the predictors. A GLM with fixed and random effects is called a generalized linear mixed model, or GLMM.

An important application of GLMMs is to item-response models, which are widely used in educational and psychological assessment. The best-known

item-response model is the Rasch model, illustrated here. Suppose 30 students are to take the same 10-question quiz, and let p_{ij} denote the probability of a correct answer by the i th student to the j th question. The Rasch model posits that

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = b_i - \beta_j,$$

where b_i represents the i th student’s ability, a random effect, and β_j represents the difficulty of the j th question, a fixed effect.

Computing for GLMMs is often challenging and requires high-speed computers and sophisticated algorithms. Indeed, the development of such algorithms is currently an active area of research among statisticians.

Other Software

Although we have used the R software in our examples, many other software packages exist. For example, the MIXED procedure in SAS is a popular alternative to the R “lme” package. S-PLUS is a commercially available package similar to R—the R commands given above will work in S-PLUS as well.

—Ronald Neath and Galin L. Jones

See also Logistic Regression Analysis; Longitudinal/Repeated Measures Data; Multivariate Normal Distribution; Rasch Measurement Model

Further Reading

- Booth, J., Hobert, J., & Jank, W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1, 333–349.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. (1996). *SAS system for mixed models*. SAS Institute.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.
- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15–32.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer-Verlag.

NLME: Software for mixed-effects models: <http://cm.bell-labs.com/stat/NLME> (also the home page for the Pinheiro and Bates reference given above, which was the source of the examples included in this entry)
 Statistical software package R: <http://www.r-project.org>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Blitstein, J. L., Murray, D. M., Lytle, L. A., Birnbaum, A. S., & Perry, C. L. (2005). Predictors of violent behavior in an early adolescent cohort: Similarities and differences across genders. *Health Education & Behavior*, 32(2), 175–194.

The authors assessed a cohort of 2,335 students from the Minneapolis, Minnesota, area to identify predictors of violent behavior and to determine whether the predictors varied by gender. The sample was 76% White; boys and girls were equally represented. The majority lived with two parents. A measure of violent behavior collected at the end of the eighth-grade year was entered into Poisson regression against baseline data collected at the beginning of the seventh-grade year. The analysis followed a general linear **mixed model** where there are multiple sources of random variation. Predictors of violent behavior influencing both boys and girls included depressive symptoms, perceived invulnerability to negative future events, paternal nonauthoritative behavior, and drinking alcohol. Additional predictors of violent behavior specific to girls included both risk and protective factors.

MIXTURE MODELS

This entry discusses statistical models involving mixture distributions. As well as being useful in identifying and describing subpopulations within a mixed population, mixture models are useful data analytic tools, providing flexible families of distributions to fit to unusually shaped data. Theoretical advances in the past 30 years, as well as advances in computing technology, have led to the wide use of mixture models in applications as varied as ecology, machine learning, genetics, medical research, psychology, reliability, and survival analysis.

Suppose that $\mathcal{F} = \{F_\theta : \theta \in S\}$ is a parametric family of distributions on a sample space X , and let Q denote a probability distribution defined on the parameter space S . The distribution

$$F_Q = \int F_\theta dQ(\theta)$$

is a mixture distribution. An observation X drawn from F_Q can be thought of as being obtained in a two-step procedure: First a random Θ is drawn from the distribution Q and then conditional on $\Theta = \theta$, X is drawn from the distribution F_θ . Suppose we have a random sample $X_1, \dots, X_n$ from F_Q . We can view this as a missing data problem in that the “full data” consist of pairs $(X_1, \Theta_1), \dots, (X_n, \Theta_n)$, with $\Theta_i \sim Q$ and $X_i | \Theta_i = \theta \sim F_\theta$ but then only the first member X_i of each pair is observed; the labels Θ_i are hidden.

If the distribution Q is discrete with a finite number k of mass points $\theta_1, \dots, \theta_k$, then we can write

$$F_Q = \sum_{j=1}^k q_j F_{\theta_j},$$

where $q_j = Q\{\theta_j\}$. The distribution F_Q is called a finite mixture distribution, the distributions F_θ are the component distributions, and the q_j are the component weights.

There are several reasons why mixture distributions, and in particular finite mixture distributions, are of interest. First, there are many applications where the mechanism generating the data is truly of a mixture form; we sample from a population that we know or suspect is made up of several relatively homogeneous subpopulations, in each of which the data of interest have the component distributions. We may wish to draw inferences, based on such a sample, relating to certain characteristics of the component subpopulations (parameters θ_j) or the relative proportions (parameters q_j) of the population in each subpopulation, or both. Even the precise number of subpopulations may be unknown to us. An example is a population of fish, where the subpopulations are the yearly spawnings. Interest may focus on the relative abundances of each spawning, an unusually low

proportion possibly corresponding to unfavorable conditions one year.

Second, even when there is no a priori reason to anticipate a mixture distribution, families of mixture distributions, in particular finite mixtures, provide us with particularly flexible families of probability distributions and densities that can be used to fit to unusually shaped (skewed, long-tailed, multimodal) data that would be difficult to describe otherwise with a more conventional parametric family of densities. Also, such a fit is often comparable in flexibility to a fully nonparametric estimate but structurally simpler, and often requires less subjective input, for example, in terms of choosing smoothing parameters. As another example, it has been shown that the very skewed log-normal density often can be well approximated by a two- or three-component mixture of normals, each with possibly different means and variances.

Formal interest in finite mixtures dates back to at least Karl Pearson's laborious method-of-moments fitting of a two-component normal mixture to data on physical dimensions of crabs in the late 19th century. The mathematical difficulties inherent in fitting mixtures have been greatly eased with the advent of the Expectation-Minimization (EM) algorithm in the 1970s. This algorithm yields an iterative method for computing maximum likelihood estimates (or very accurate approximations thereof) in a general missing-data situation. As mentioned above, mixtures have a natural missing-data interpretation and so the EM algorithm, together with improved computing technology, has made the task of fitting mixtures models to data much easier, leading to a renewal of interest in them.

Fitting Finite Mixtures Using Maximum Likelihood

The EM algorithm generates a sequence of parameter estimates, each of which is guaranteed to give a larger likelihood than its predecessor. It can be used whenever the original log-likelihood $\log f_X(x; \theta)$ is difficult to maximize over θ for given x , but $f_X(x; \theta)$ can be expressed as the marginal distribution of X in a pair

(X, J) whose corresponding log-likelihood $\log f_{XJ}(x, j; \theta)$ is easier to maximize over θ for a given x and j . Given a "current estimate" θ_0 , the next in the sequence, θ_1 , is defined as the maximizer of the EM-log-likelihood $\ell_{EM}(\theta; x)$, which is defined as the conditional expectation of $\log f_{XJ}(x, J; \theta)$ over the "missing data" J given $X = x$ computed under θ_0 ; that is,

$$\ell_{EM}(\theta; x) = E \log f(x, J; \theta) \text{ where } J \text{ has density } f_{J|X}(j | x; \theta_0) = f_{XJ}(x, j; \theta_0) / f_X(x; \theta_0).$$

It is guaranteed that $\log f_X(x; \theta_1) \geq \log f_X(x; \theta_0)$.

If we wish to fit a finite mixture

$$f(x; Q) = \sum_{j=1}^k q_j f(x; \theta_j)$$

where the number of components k is known, the EM algorithm works in almost the same way for either one or both of the q_j s or θ_j s unknown. We regard the x_i s as the observed first members of random pairs $(X_1, J_1), \dots, (X_n, J_n)$, but the J_i s are unobserved. We can write the full data log-likelihood as

$$\sum_{i=1}^n \sum_{j=1}^k 1\{J_i = j\} \{\log q_j + \log f(x_i; \theta_j)\}$$

(here $q_j = P\{J_i = j\}$). We now outline how to go from an initial set of estimates $q_{01}, \dots, q_{0k}, \theta_{01}, \dots, \theta_{0k}$ to the next in the EM sequence $q_{11}, \dots, q_{1k}, \theta_{11}, \dots, \theta_{1k}$. If some of these values are known, then they of course remain unchanged. The first step is to compute the posterior probabilities

$$\pi_{ji} = P\{J_i = j | X_i = x_i\} \text{ computed under the } q_{0j}\text{s and } \theta_{0j}\text{s}$$

$$= \frac{q_{0j} f(x_i; \theta_{0j})}{\sum_{j=1}^k q_{0j} f(x_i; \theta_{0j})}.$$

The EM-log-likelihood is then obtained by replacing the $1\{J_i = j\}$ s in the full data log-likelihood with the π_{ji} s; note that the EM-log-likelihood thus obtained separates into a term involving the q_j s only and one involving the θ_j s only.

If the q_j s are unknown, we maximize

$$\sum_{j=1}^k \log q_j \left\{ \sum_{i=1}^n \pi_{j|i} \right\}$$

with respect to the q_j s; this is maximized at

$$q_{1j} = n^{-1} \sum_{i=1}^n \pi_{j|i},$$

simply the averages of the posterior probabilities over the data.

If the θ_j s are unknown, we maximize

$$\sum_{j=1}^k \sum_{i=1}^n \pi_{j|i} \log f(x_i; \theta_j)$$

with respect to the θ_j s. Differentiating with respect to each θ_j and setting to zero yields k weighted score equations:

$$\sum_{i=1}^n \pi_{j|i} \frac{\partial \log f(x_i; \theta_j)}{\partial \theta_j} = 0.$$

In many common models, these are easily solved. For example, in one-parameter exponential families of the form $f(x; \theta) = e^{\theta x - K(\theta)} f_0(x)$ (e.g., normal with known variance, Poisson), let $\hat{\theta}(t)$ be that value of θ that solves $K'(\theta) = t$. Then, for each j , one can explicitly find the EM update as

$$\theta_{j1} = \hat{\theta} \left(\frac{\sum_{i=1}^n \pi_{j|i} x_i}{\sum_{i=1}^n \pi_{j|i}} \right),$$

a known function of a $\pi_{j|i}$ -weighted average of the x_i s.

Further Inferences

Once the model has been fitted, further inferences may consist of confidence intervals for or hypothesis tests concerning the component parameters θ_j and/or the mixing proportions q_j . When the model is correctly specified (i.e., there really are k components and all the q_j s are positive), the parameter estimates behave more or less in a standard fashion: They are

asymptotically normal with an estimable covariance matrix, subject to the component densities $f(x; \theta_j)$ being suitably regular. Hence, confidence regions can be computed in a standard fashion, bearing in mind the restrictions on the q_j s: They are nonnegative and add to 1. In addition, one should be aware that when the weights q_j are small or the parameters θ_j for two or more groups are similar, there is a sharp loss of estimating efficiency as well as a good reason to be doubtful of accuracy of asymptotic approximations. This occurs because of the near loss of identifiability of the parameters near the boundaries of the parameter space.

Hypothesis tests are perhaps not so standard, at least not for tests concerning the q_j s. If one wishes to test whether an estimate $\hat{q}_j$ is significantly different from zero, the nonnegativity constraints have a significant impact, at least when it comes to using large-sample χ^2 approximations to the p values. Because such a hypothesis constrains a parameter to be on the boundary of the parameter space, the asymptotic distribution of twice the log-likelihood ratio will be a mixture of χ^2 distributions rather than a pure χ^2 , assuming the model is otherwise suitably regular. In such a case, a parametric bootstrap approach can be used to obtain an approximate p value.

An Unknown Number of Components, or Completely Unknown Q

If the number of components of a putatively finite mixture is unknown, we are essentially on the same footing as knowing absolutely nothing about Q , for reasons we now explain.

For any given data set $x_1, \dots, x_n$ with $d \leq n$ distinct x_i s and any prespecified Q , no matter whether it is discrete or continuous, so long as the likelihoods $f(x; \theta)$ are bounded in θ , we can find a discrete $\tilde{Q}$ with $m \leq d$ support points such that Q and $\tilde{Q}$ provide exactly the same density values at the observed data. That is, for any mixing distribution Q , there is a possibly different $\tilde{Q}$ yielding a finite mixture such that Q and $\tilde{Q}$ cannot be distinguished, at least in terms of the data $x_1, \dots, x_n$. So it suffices to restrict attention to such $\tilde{Q}$ s.

An implication of this, when the likelihoods are bounded in θ , is that the maximum likelihood estimate of Q over all distributions, which we denote by $\hat{Q}$, exists and is finite with at most d (the number of distinct x_i s) support points. So we need never leave the realm of finite mixtures in this setting.

This is not to say, however, that an estimate of an unknown k is readily available. The number of components in $\hat{Q}$ may be an overestimate in that some support points (respectively mixing proportions) may be so close together (small) that combining them into a single point (removing them) hardly decreases the likelihood. This and other issues related to trying to infer something about the number of components in a mixture, like hypothesis tests concerning k , are difficult problems. Some problems are still open, and others have solutions that are possibly too complex to be useful.

The Nonparametric Estimate of Q

When the estimate $\hat{Q}$ discussed above exists, it is discrete with at most d support points. Hence, a strategy for computing it is to try to fit a finite mixture with d components using the EM algorithm. In many situations, this yields a sensible result. More sophisticated algorithms exist, however, that are related to the following gradient function characterization.

The gradient function

$$D_Q(\theta) = \sum_{i=1}^n \left[\frac{f(x_i; \theta)}{f(x_i; Q)} - 1 \right]$$

measures the rate of increase in the log-likelihood if we remove a small amount of weight from the mixing distribution Q and put it at the point θ . Hence, for a candidate estimate Q , if for some θ we have $D_Q(\theta) > 0$, we know that we can increase the log-likelihood by putting some weight at θ .

In light of this, the following result is not surprising: If the nonparametric maximum likelihood estimate $\hat{Q}$ exists, then $D_{\hat{Q}}(\theta) \leq 0$ for all θ , and the support points of $\hat{Q}$ are included in the set of values θ where $D_{\hat{Q}}(\theta) = 0$. The fact that $D_{\hat{Q}}(\theta) > 0$ for no θ makes sense; moving mass around from $\hat{Q}$ to any other θ cannot increase the likelihood.

The nonparametric version of the mixture model falls into the class of convex models, a subject with its own independent literature. Often, convex models can be written as mixture models. For example, a distribution function that is concave on the positive half-line can also be written as a nonparametric mixture of the form $\int f(x; \theta) dQ(\theta)$ with component density $f(x; \theta) = 1\{0 < x < \theta\}/\theta$. One can deduce that the nonparametric likelihood estimator is the least concave majorant of the empirical distribution function using the above gradient characterization.

—Bruce Lindsay and Michael Stewart

See also Likelihood Ratio Test

Further Reading

- Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, 11, 36–53.
- Liu, A. (2002). Efficient estimation of two seemingly unrelated regression equations. *Journal of Multivariate Analysis*, 82, 445–456.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

MIXTURES OF EXPERTS

It is often the case that the analysis of the values of a set of random variables becomes simpler if one posits that these variables are related to another set of variables, called latent variables, whose values are unobserved. Consider, for example, the two-dimensional data in Figure 1. This data set is complicated in the sense that it is multimodal and, thus, cannot be summarized by a standard distribution such as those comprising the exponential family. A way of simplifying the analysis is by assuming that the distribution of these data is a combination of two simple distributions, namely, two Gaussian distributions. Each data item was generated as follows. First, a value for a latent variable was sampled from a Bernoulli distribution. Next, if the latent variable was set to 0, then the data item was sampled from the first Gaussian distribution

(e.g., with mean vector $[3 \ 3]^T$); if the latent variable was set to 1, then the data item was sampled from the second Gaussian distribution (with mean vector $[7 \ 7]^T$). Although the value of the latent variable is not observed, it is easy to use Bayes's rule to compute the distribution of the variable given the data item.

This model is a latent variable model known as a mixture model. Mixture models provide a principled way of combining two or more simple distributions (e.g., unimodal distributions such as Gaussian distributions) into a single complicated (e.g., multimodal) distribution. As this example illustrates, mixture models are “piecewise estimators” in the sense that different components are used to summarize different subsets of the data. The subsets do not, however, have hard boundaries; as discussed below, a data item might be a member of multiple subsets simultaneously.

Mixtures-of-experts (ME) models are an extension of mixture models. They differ from conventional mixture models in that their mixture components are conditional probability distributions. Consequently, they are suitable for summarizing data sets in which the distribution of output or response variables depends on the values of input or covariate variables. Such data sets arise in the context of regression or classification tasks, for example.

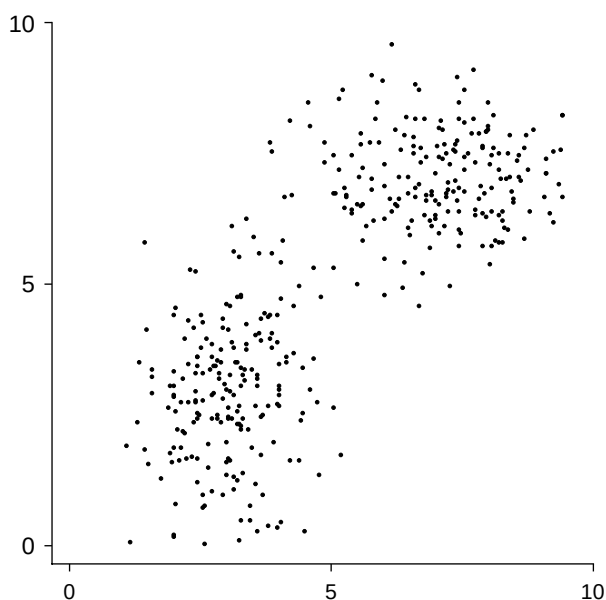


Figure 1 Two-Dimensional Data to Be Summarized

ME models perform tasks using a “divide and conquer” strategy—complex tasks are decomposed into simpler subtasks. ME models can be characterized as fitting piecewise models to the data. The data are assumed to form a countable set of paired variables $X = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^T$, where $\mathbf{x}$ is a vector of explanatory variables, also referred to as covariates, and $\mathbf{y}$ is a vector of responses. ME models divide the covariate space, meaning the space of all possible values of the explanatory variables, into regions, and then they fit simple surfaces to the data that fall in each region. Unlike many other piecewise approximators, these models use regions that are not disjoint. The regions have “soft” boundaries, meaning that data points may lie simultaneously in multiple regions. In addition, the boundaries between regions are themselves simple parameterized surfaces whose parameter values are estimated from the data.

ME models combine properties of generalized linear models with those of mixture models. Like generalized linear models, they are used to model the relationship between a set of covariate and a set of response variables. Unlike standard generalized linear models, however, they assume that the conditional distribution of the responses (given the covariates) is a finite mixture distribution. Because ME models assume a finite mixture distribution, they provide a motivated alternative to nonparametric models and provide a richer class of distributions than standard generalized linear models.

Mixtures-of-Experts Model

To define the mixtures-of-experts model, suppose that the process generating the data is decomposable into a set of subprocesses defined on possibly overlapping regions of the covariate space. For each data item, a subprocess is selected, conditional on the covariate $\mathbf{x}^{(i)}$, and the selected subprocess maps $\mathbf{x}^{(i)}$ to the response $\mathbf{y}^{(i)}$. Specifically, for each covariate $\mathbf{x}^{(i)}$,

- a label i is selected from a multinomial distribution with probability $P(i | \mathbf{x}^{(i)}, V)$, where $V = [\mathbf{v}_1, \dots, \mathbf{v}_I]$ is the matrix of parameters underlying the multinomial distribution

- a response $\mathbf{y}^{(t)}$ is generated with probability $P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, U_i, \Phi_i)$, where U_i is a parameter matrix and Φ_i represents other parameters, for $i = 1, \dots, I$.

To explicitly relate this approach to the generalized linear models framework, it is supposed that the conditional probability distribution $P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, U_i, \Phi_i)$ is a member of the exponential family of distributions. The expected conditional value of the response $\mathbf{y}^{(t)}$, denoted $\boldsymbol{\mu}_i^{(t)}$, is defined to be a generalized linear function f of $\mathbf{x}^{(t)}$, as well as the parameter matrix U_i . The quantities $\eta_i = f^{-1}(\boldsymbol{\mu}_i)$ and Φ_i are, respectively, the natural parameter and dispersion parameter of the response's distribution $P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, U_i, \Phi_i)$. The total probability of generating the response $\mathbf{y}^{(t)}$ from $\mathbf{x}^{(t)}$ is given by the mixture density

$$P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, \Theta) = \sum_i P(i | \mathbf{x}^{(t)}, V) P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, U_i, \Phi_i), \quad (1)$$

where $\Theta = [\mathbf{v}_1, \dots, \mathbf{v}_p, U_1, \dots, U_p, \Phi_1, \dots, \Phi_p]^T$ is the matrix of the parameters. Assuming independently distributed responses, the total probability of the data set X is the product of T such densities, with likelihood

$$L(\Theta | X) = \prod_t \sum_i P(i | \mathbf{x}^{(t)}, V) P(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, U_i, \Phi_i). \quad (2)$$

Figure 2 presents a graphical representation of the mixtures of experts model. The model consists of n modules referred to as *expert networks*. These networks approximate the data within each region of the covariate space: Expert network i maps its input, the covariate vector $\mathbf{x}$, to an output vector $\boldsymbol{\mu}_i$. It is supposed that different expert networks are appropriate in different regions of the covariate space. Thus, the model requires a module, referred to as a *gating network*, that identifies for any covariate $\mathbf{x}$ the expert or mixture of experts whose output is most likely to approximate the corresponding response vector $\mathbf{y}$. The gating network outputs are a set of scalar coefficients g_i that appropriately weights the contributions of each expert. For each covariate $\mathbf{x}$, these coefficients sum to one and are constrained to be

nonnegative. The total output of the mixtures of experts model, given by

$$\boldsymbol{\mu} = \sum_{i=1}^n g_i \boldsymbol{\mu}_i, \quad (3)$$

is a combination of the expert outputs for each $\mathbf{x}$.

From the point of view of mixture modeling, we identify the gating network with the selection of a particular subprocess. That is, the gating outputs g_i are interpreted as the covariate dependent, multinomial probabilities of selecting subprocess i . Different expert networks are identified with different subprocesses, and each expert models the covariate dependent distributions corresponding to its particular subprocess.

The expert networks map their inputs to their outputs in a two-stage process. During the first stage, each expert multiplies the covariate vector $\mathbf{x}$ by a matrix of parameters. (The vector $\mathbf{x}$ is required to include a fixed component of one to allow for an intercept term.) For expert i , the matrix is denoted as U_i and the resulting vector is denoted as $\boldsymbol{\eta}_i$, where

$$\boldsymbol{\eta}_i = U_i \mathbf{x}. \quad (4)$$

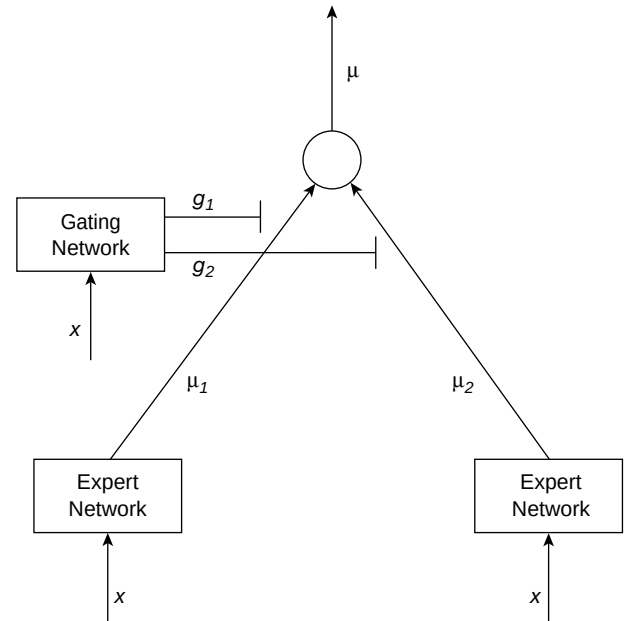


Figure 2 A Mixture-of-Experts Model

During the second stage, η_i is mapped to the expert output μ_i by a monotonic, continuous nonlinear function f .

The selection of the nonlinear function f is based on the nature of the problem. For regression problems, f may be taken as the identity function (i.e., the experts are linear) and the probabilistic component of the model may be Gaussian. In this case, the likelihood is a mixture of Gaussians:

$$L(\Theta | X) = \prod_t \sum_i g_i^{(t)} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}[\mathbf{y}^{(t)} - \mu_i^{(t)}]^T \Sigma_i^{-1} [\mathbf{y}^{(t)} - \mu_i^{(t)}]}. \quad (5)$$

The expert output $\mu_i^{(t)}$ and expert dispersion parameter Σ_i are the mean and covariance matrix of expert i 's Gaussian distribution for response $\mathbf{y}^{(t)}$ given covariate $\mathbf{x}^{(t)}$. The output of the entire model $\mu^{(t)}$ is interpreted as the expected value of $\mathbf{y}^{(t)}$ given $\mathbf{x}^{(t)}$.

For binary classification problems, f may be the logistic function

$$\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \quad (6)$$

In this case, η_i may be interpreted as the log-odds of "success" under a Bernoulli probability model. The probabilistic component of the model is generally assumed to be the Bernoulli distribution. The likelihood is a mixture of Bernoulli densities:

$$L(\Theta | X) = \prod_t \sum_i g_i^{(t)} [\mu_i^{(t)}]^{y^{(t)}} [1 - \mu_i^{(t)}]^{1-y^{(t)}}. \quad (7)$$

The quantity $\mu_i^{(t)}$ is expert i 's conditional probability of classifying the covariate $\mathbf{x}^{(t)}$ as success, and μ is the expected success of $\mathbf{x}^{(t)}$. Other problems (e.g., multiway classification, counting, rate estimation, survival estimation) may require other choices for f . In all cases, the inverse of f is taken to be the canonical link function for the appropriate probability model.

The gating network also forms its outputs in two stages. During the linear stage, it computes the

intermediate variables ξ_i as the inner product of the covariate vector $\mathbf{x}$ and the vector of parameters $\mathbf{v}_i$:

$$\xi_i = \mathbf{v}_i^T \mathbf{x}. \quad (8)$$

The ξ_i are mapped to the gating outputs g_i during the nonlinear stage. This mapping is performed by using a generalization of the logistic function:

$$g_i = \frac{e^{\xi_i}}{\sum_j e^{\xi_j}}. \quad (9)$$

In the neural network literature, this function is known as the *softmax function*. Note that the inverse of this function is the canonical link for a multinomial response model.

Maximum likelihood estimates of an ME model's parameter values can be obtained using the Expectation-Maximization algorithm. Alternatively, Bayesian inference regarding an ME model's parameters can be performed using Markov chain Monte Carlo methods. Furthermore, the basic ME model can be extended to a hierarchical mixtures-of-experts model that divides tasks into subtasks, and subtasks into sub-subtasks. The recursive nature of this hierarchical extension makes it an efficient and appealing model for many data sets.

—Martin A. Tanner and Robert A. Jacobs

See also Poisson Distribution

Further Reading

- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79–87.
- Jacobs, R. A., Tanner, M. A., & Peng, F. (1996). Bayesian inference for hierarchical mixtures of experts with applications to regression and classification. *Statistical Methods in Medical Research*, 5, 375–390.
- Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181–214.
- Peng, F., Jacobs, R. A., & Tanner, M. A. (1996). Bayesian inference in mixtures of experts and hierarchical mixtures of experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91, 953–960.

MODE

The mode is one characteristic of a distribution. It is the most frequently occurring value in a group of values. The mode can be an informative description when one value occurs noticeably more often than other values. In addition, the mode is the only measure of average, or central tendency, that is appropriate for nominal (categorical) data.

For example, the age distribution in a group of 12 children is 3, 3, 4, 4, 4, 4, 4, 5, 6, 6, 7, and 8. To compute the mode, the frequency of all occurring values has to be calculated.

Value	Frequency
Age 3	2
Age 4	5
Age 5	1
Age 6	2
Age 7	1
Age 8	1

The value that occurs most frequently is a mode of the distribution. In this example, 4 is the modal age of the children. The mode tells us which age occurred more often than others.

In distributions of continuous variables, the mode can be computed as the midpoint of the histogram with the highest peak. It can be useful to revert to modal classes, sometimes called “bins” (e.g., a class of values between 1.1 and 2, 2.1 and 3), which can subsume the individual values 1.23, 1.76, 2.11, or 2.89. Note that this can cause ambiguity, because the choice of class boundaries can have a dramatic effect on the mode.

A distribution that has two peaks is called *bimodal*; when there are more than two peaks, it is called *multimodal*.

For the distribution to be bimodal, the two peaks do not necessarily have to be of equal height.

More About the Mode

- A histogram is an easy method to spot modes.
- If no value occurs more often than another in the group of values, any and every value can be considered a mode of the distribution.
- The mode is independent of the range and the shape of distribution of the represented values. Unlike the mean, it is not affected by extreme values because it does not take other values of the distribution into account. If three people had a score of 3 on a test, the mode is 3 regardless of the fact that other people's test scores ranged from 1 to 100 (because only one or two people had the same score).

Analysis Using SPSS

Figure 1 is a simple output using SPSS's descriptive feature.

—Susanne Hempel

See also Average; Mean; Measures of Central Tendency; Median

Frequencies

Statistics

age		
N	Valid	12
	Missing	0
Mode		4

age

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 3	2	16.7	16.7	16.7
4	5	41.7	41.7	58.3
5	1	8.3	8.3	66.7
6	2	16.7	16.7	83.3
7	1	8.3	8.3	91.7
8	1	8.3	8.3	100.0
Total	12	100.0	100.0	100.0

Figure 1 SPSS Output

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

MODERATOR VARIABLE

A moderator variable is an independent or predictor variable (e.g., Z) that interacts with another independent or predictor variable (e.g., X) in predicting scores on and accounting for variance in a dependent or predicted variable (e.g., Y). Note that the terms *independent variable* and *dependent variable* are appropriate only in the case of experimental research. In nonexperimental research, the appropriate, analogous terms are *predictor variable* and *predicted variable*.

Moderator variables (referred to hereinafter as moderators) imply conditional relations. That is, the strength and/or form (e.g., linear, quadratic) of the relation between X and Y varies as a function of the value of the moderator, Z . There are many examples of moderators in both theory and research in such disciplines as psychology, sociology, management, education, political science, biology, epidemiology, and medicine. For instance, in industrial and organizational psychology, extant theory specifies that the relation between ability and performance is moderated by motivation. The greater the level of motivation, the stronger the relation between ability and performance.

By studying moderators, researchers learn about how relations between variables of interest vary across levels of the moderator. Interestingly, it has been argued that the amount of progress in any discipline can be indexed by the degree to which its theory and research have considered the role of moderators.

An Important Distinction

A review of the literature reveals that moderators are often confused with mediator variables (referred to below as mediators). In contrast to the just noted role

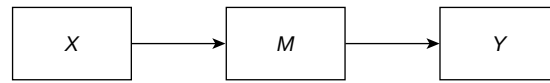


Figure 1 Illustration of a Mediating Effect

played by moderators, mediators transmit the effects of independent variables to dependent variables. This is illustrated in the causal chain shown in Figure 1, which depicts M as a mediator of the relation between X and Y . There are many examples of mediation. In psychology, for instance, stress (M) has been shown to mediate the relation between stressors (X) and strain (Y).

Strategies for Detecting Moderators

Over the years, a number of strategies have been proposed for the detection and description of moderators. Unfortunately, many strategies are unsound (e.g., single group validity). In addition, others (e.g., comparing correlation coefficients for artificially created subgroups) have lower levels of statistical power than others (e.g., moderated multiple regression). In the interest of brevity, only three strategies are described here: (a) the analysis of variance based test for an interaction, (b) the multiple regression based test for an interaction, and (c) the test for the homogeneity of correlation coefficients.

The Analysis of Variance Strategy

A commonly used technique for testing interaction effects in *experimental* research is analysis of variance (ANOVA). Assuming a 2×2 experimental design involving independent variables A ($a_1, a_2, \dots, a_j$) and B ($b_1, b_2, \dots, b_k$), ANOVA tests for an interaction by determining if the $A \times B$ effect explains variance over and above the additive effects of A and B . The respective population and sample effect models for a two-way ANOVA are as follows:

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk}$$

$$y_{ijk} = y + a_j + b_k + (ab)_{jk} + e_{ijk}$$

In terms of the population model, a main effect of A is implied by the rejection of the null hypothesis of $\alpha_j = 0$, a main effect of B is indicated by the rejection of the null hypothesis of $\beta_j = 0$, and an interaction effect is signaled by the rejection of the null hypothesis of $(\alpha\beta)_{jk} = 0$.

Figure 2 shows a plot of cell means for a 2×2 experimental design in which (a) there are main effects for both A and B , and (b) there is an $A \times B$ (interaction) effect. In terms of moderation, the effect of A on Y is stronger at B_2 than B_1 , as is suggested by the fact that the line at B_2 has a steeper slope than the line at B_1 .

The Moderated Regression Strategy

In nonexperimental research, moderated multiple regression (MMR) is an appropriate and commonly used strategy for testing moderating effects. Note that MMR also may be used for the same purposes when a researcher is dealing with data from an experimental study. It merits adding that when MMR is used for the analysis of data from nonexperimental research, the term *effects* is used in a statistical (model testing) sense, not to imply that causal relations are being tested.

The MMR strategy relies on the test of a linear model having terms for hypothesized main (X , Z) and interactive ($X \times Z$) effects. A sample-based regression

equation for a model involving predictors X and Z (the hypothesized moderator) and a product term that carries information about the interaction (XZ) is as follows:

$$\hat{Y} = b_0 + b_1X + b_2Z + b_3XZ.$$

It is important that X and Z scores be centered (e.g., standardized) prior to the computation of the $X \times Z$ product term. Moderation is signaled by the rejection of the null hypothesis of $\beta_3 = 0$. This hypothesis is tested using the sample based regression coefficient for b_3 in the just noted equation.

The moderator (Z) can be nominal (k groups) or continuous. When nominal, the moderator is represented in the regression equation uses by $k - 1$ dummy variables. For example, if the moderator is sex (e.g., female, male), Z is represented in the regression equation by a single dummy variable. When Z is continuous, it is represented in the regression equation by a single variable.

Information produced by an MMR analysis can be used to show the nature of the moderating effect. Figure 3 illustrates the regression lines for a discrete moderator. As can be seen in the figure, the slope of the regression line is steeper when Z equals 1 than when it equals zero. Figure 4 depicts the regression surface for a continuous moderator. Regression lines are shown for four illustrative levels of Z . Note that the greater the value of Z , the steeper the slope of the Y on X regression.

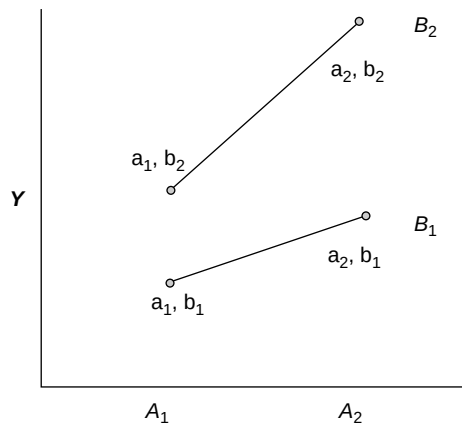


Figure 2 Interaction Effect for a 2×2 Experimental Study

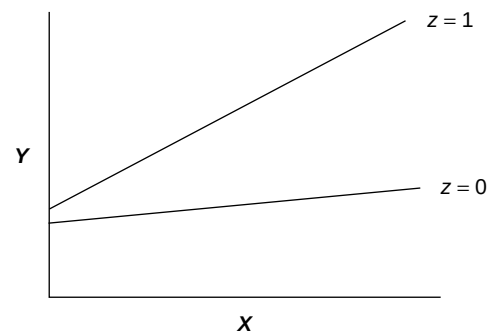


Figure 3 Moderating Effect for a Discrete Moderator in a Nonexperimental Study

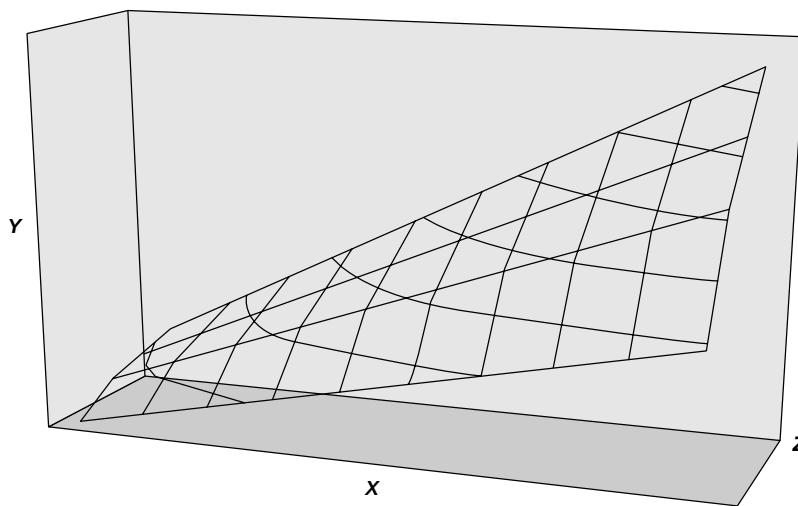


Figure 4 Moderating Effect for a Continuous Moderator in a Nonexperimental Study

The Homogeneity of Correlation Coefficients Strategy

When a moderator is nominally scaled naturally (e.g., fields of psychology, religious sects), moderating effects can be assessed by testing the equality of XY correlation coefficients for k groups. The null hypothesis is as follows:

$$H_0 : \rho_1 = \rho_2 = \rho_3 = \dots = \rho_k.$$

This hypothesis is tested using sample estimates of the k population correlation coefficients. A moderating effect is inferred when the test allows for the rejection of the null hypothesis.

Some Other Considerations

This entry provides a brief summary of some major issues surrounding the detection of moderators and the description of moderating effects. There are a host of other issues surrounding moderators. For example, the ability to detect moderators is a function of such factors as sample size, strength of moderating effect, reliability of measures of the moderator, range restriction on the moderator, and the interaction of several of these factors. Readers are encouraged to learn more

about these issues before doing research that involves moderators.

—Eugene F. Stone-Romero

See also Independent Variable; Tests of Mediating Effects

Further Reading

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Hall, J. M., & Rosenthal, R. (1991). Testing for moderator variables in meta analysis: Issues and methods. *Communication Monographs*, 58, 437–448.
- Stone, E. F. (1988). Moderator variables in research: A review and analysis of conceptual and methodological issues. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 6, pp. 191–229). Greenwich, CT: JAI.
- Stone-Romero, E. F. (2007). Nonexperimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F. (2007). Quasi-experimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F. (2007). Randomized experimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F., & Liakhovitski, D. (2002). Strategies for detecting moderator variables: A review of theory and research. *Research in Personnel and Human Resources Management*, 21, 333–372.
- Zedeck, S. (1971). Problems with the use of “moderator” variables. *Psychological Bulletin*, 76, 295–310.

MONTE CARLO METHODS

The probability of an obverse surfacing was estimated by George Louis Leclerc, Comte de Buffon (1707–1788), by flipping a coin 4,040 times. He obtained 2,048 heads, which produced an estimate of $\frac{2048}{4040} = 50.69\%$. As part of a genetics experiment, Walter

Frank Raphael Weldon (1860–1906) tossed 12 dice at the same time, recorded the results, and repeated the process 26,306 times. (Due to the tediousness and error-prone method of physically counting the results, Karl Pearson (1857–1936) supported the veracity of the outcome with the χ^2 test.) William Sealy Gosset used repeated sampling with slips of stiff paper to find the distribution of the correlation coefficient and to support the development of the t distribution.

These experiments are examples of the Monte Carlo method. It is a resampling approximation technique. The name is derived from the casinos of the principality of Monte Carlo.

The utility and accuracy of the Monte Carlo method was greatly enhanced with the advent of the computer and software. Stanislaw Marcin Ulam (1909–1984) wrote a Monte Carlo computer simulation of the solitaire card game in 1946. In a more important application, along with his boss John (János) von Neumann (1903–1957) and Enrico Fermi (1901–1954), Ulam estimated the eigenvalues for the Schrödinger equation (Erwin Schrödinger, 1887–1961) using Monte Carlo methods. Subsequently, he developed a Monte Carlo computer simulation of random neutron diffusion in fissile material to construct dampers and shields for the atomic bomb as part of the Manhattan Project.

Pseudo Random Number Generators

Initially, the basis of the Monte Carlo method was the use of uniform pseudo random numbers on the interval $[0,1]$. Today, Monte Carlo methods (plural) apply to the use of any pseudo random number generator, such as variates obtained from the exponential distribution, or repeated sampling from large, real data sets.

Programming Environment

Since its commercial release by the IBM Corporation in 1957, FORTRAN (FORMula TRANslator) remains the fastest high-level programming language for Monte Carlo simulation work. This is because execution time is an essential component for realistic, applied problems. Higher level programming environments,

such as SAS IML, S+, and Lucent Technology's R (which is available at no cost from <http://www.rproject.org/>), are serviceable for simple Monte Carlo simulations and classroom demonstrations.

Simulating Tossing a Die

A Monte Carlo computer solution easily simulates the tossing of a fair die with a variate drawn from a uniform $[0,1]$ pseudo random number generator and a table of assigned outcomes, such as indicated in Table 1. For example, if the value obtained from the generator is .1770, it simulates the throwing of a fair die and obtaining two spots. Counting the results of repeating this simulation is tremendously faster and more accurate than physically tossing a die.

Table 1 Simulation of a Fair Die Using Uniform Variates on the Interval $[0,1]$

<i>Outcome Assignment</i>	
.0000 – .1666	1 spot
.1667 – .3333	2 spots
.3334 – .5000	3 spots
.5001 – .6666	4 spots
.6667 – .8333	5 spots
.8334 – 1.000	6 spots

Source: Sawilowsky (2003, p. 219).

Estimating the Area of a Regular Figure

Consider the area (A) of a well-defined geometric shape, such as that in the shaded area depicted in Figure 1. The area of interest is bounded by the two equations $f(x) = x$ and $g(x) = x^2$ over the interval $(0, 1)$.

The shaded area can be estimated via the Monte Carlo method by repeatedly drawing pairs $(x$ and $y)$ of pseudo random uniform numbers. If $x > y$, the coordinate pair (x, y) is below the line $f(x) = x$. If $x^2 < y$, then (x, y) is above $g(x) = x^2$. If both conditions are true, the ordered pair (x, y) falls within the shaded region. The number of times the paired coordinate (x, y) falls in the shaded region, divided by the total number of

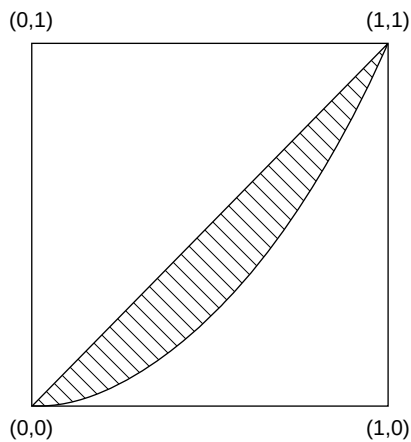


Figure 1 Estimating Area

Source: Sawilowsky & Fahoome (2003, p. 120).

iterations, multiplied by the total area (which in this case is 1.0), is an estimate of the area of the shaded region.

By inspection, $f(x)$ bisects the unit square, meaning that the area above or below the diagonal line $f(x) = x$ is .5. The Monte Carlo results indicate the area below $g(x) = x^2 \approx .33$, and the shaded area $(A) \approx .17$.

Estimating the Area of an Irregular Figure

The shaded area in the previous example is obtained easily and precisely through elementary calculus. The Monte Carlo method is useful, however, in determining the area of an irregular figure where the calculus either proves difficult or is intractable. For example, consider a unit square with an inscribed arbitrary figure with a curvilinear boundary. The figure need not be contiguous.

Draw a pair (x and y) of pseudo random uniform $[0,1]$ numbers to represent a coordinate pair (x,y) . If it falls within the inscribed figure, increase a counter (N'). Repeat the process N times. The ratio $\frac{N'}{N}$ is the Monte Carlo estimate of the area of the inscribed figure.

Estimating Irrational Numbers

Monte Carlo methods can be used to estimate irrational numbers, such as π . In 1777, Buffon posed a problem pertaining to a surface marked with parallel

lines, with the distance d from each line. If a needle of length $L < d$ is dropped at random on the surface, the probability that it will cross a line is

$$\int_0^\pi \int_0^{L \sin \theta} \frac{1}{d\pi} dA d\theta,$$

where θ is the angle between the needle and the lines. This leads to a Monte Carlo estimate of π as

$$\pi = \frac{2nL}{Md}.$$

Buffon dropped a needle 2,000 times on the above-described surface. Measurements yielded an estimate for $\pi = 3.143$.

An estimate of π may be obtained more simply via Monte Carlo methods on a computer by considering Figure 2. It represents the upper right quadrant of a Cartesian plane where a sector of a circle with the center at $(0, 0)$ has been inscribed in a unit square. Obtain two pseudo random uniform variates on the interval $(0, 1)$, letting the first value represent the x coordinate and the second value represent the y coordinate (x, y) .

The radius of the circle is 1. The length L of the line from the origin to the coordinate pair is determined by

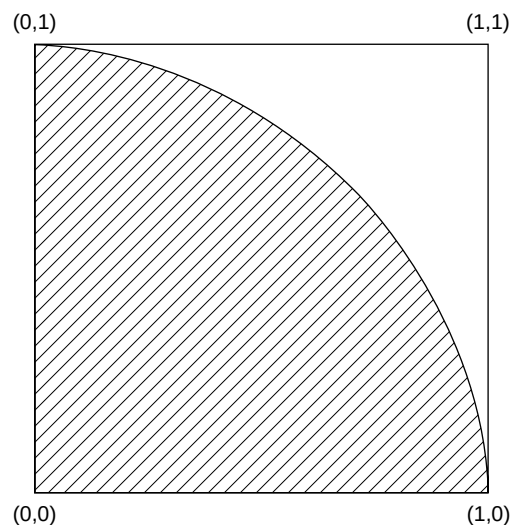


Figure 2 Estimating π

Source: Sawilowsky & Fahoome (2003, p. 123).

the Pythagorean theorem $L = \sqrt{x^2 + y^2}$. If the length is less than the radius of 1, then the point lies within the circle. Otherwise, it is outside the circle but within the square.

The area of a circle, $A = \pi r^2$, reduces to $A = \pi$ when the radius is 1. Because the sector represents 1/4 of a circle, π can be estimated by multiplying this area by 4. As the number of repetitions (e.g., pairs of pseudo random uniform variates) increases, the accuracy of the estimate of π increases.

Monte Carlo Option in Statistics Software Packages

Modern statistical practice has been heavily affected by the availability of high-speed computers that permit hypothesis tests based on permutation methods. However, many statistical problems remain beyond current capabilities. An alternative is to conduct an approximate permutation statistic, which is based on a specific subset of all possible permutations.

Approximate permutation methods require keeping a record of permutations that have been used in the analysis, which can be as burdensome as performing all possible permutations. An estimation procedure that is easily implemented is a Monte Carlo option, where a subset of permutations are randomly selected. Although there may be some overlapping of selected permutations, as the number of repetitions increases, the impact of duplicated permutations becomes trivial.

Many statistical software packages have a Monte Carlo option, such as StatXact, SAS, and SPSS. Subroutines (called macros) can be written for Minitab, and well as for spreadsheet software such as Excel. For example, to compute a stratified $2 \times C$ Wilcoxon Mann Whitney test using the Monte Carlo option in StatXact, select Statistics | $2 \times C$ Tables | Exact using Monte Carlo | OK.

Accuracy of Monte Carlo

Reconsider the problem of estimating the area (A) of an irregular figure. As $N \rightarrow \infty \frac{N'}{N} \rightarrow A$. Because

$$\text{error} \approx \sqrt{\frac{A(1-A)}{N}},$$

N must increase by a factor of 10^2 to obtain a new significant digit. Sobol recommended the Monte Carlo method where the desired level of accuracy is no less than 5%–10%, because an increase in N by a factor of 10^2 was somewhat daunting in the early years of computing. However, increasing the number of repetitions is trivial with modern computers and software, permitting nearly any desired level of accuracy.

In general, stochastic (and nonstochastic) Monte Carlo studies can be reduced to

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^n f(U^{[i]}),$$

where the random variable $f(U)$ is the mean, and $U = (U^{[1]}, \dots, U^{[n]})$ is an n dimensional uniform $[0,1]$ variate. The standard error σ_n decreases with the square root of the sample size n :

$$\sigma_n = \frac{\sigma}{\sqrt{n}}.$$

Thus, the utility of Monte Carlo methods becomes evident in considering a family of all numerical methods in m dimensional space that are functions of r point estimations (repetitions). The absolute errors decrease at the rate of only $r^{-1/m}$, whereas the Monte Carlo method's absolute error decreases at the rate of $r^{-1/2}$.

Monte Carlo Study Versus Monte Carlo Simulation

Sawilowsky differentiated between a Monte Carlo study and a Monte Carlo simulation. Determining that the expected value $E[x]$ of a uniform $[0,1]$ continuous variable is $\int_0^1 f(x)dx = .5$ via a Monte Carlo study can be accomplished by drawing N variates from the uniform distribution, summing them, and dividing the total by N . This does not necessarily simulate any real phenomenon. Because the probability $P[x]$ of two spots surfacing on a die is $\frac{1}{6}$, drawing variates from the uniform distribution and assigning spots according to Table 1 is a Monte Carlo simulation of throwing a die.

Four important characteristics of high-quality Monte Carlo study are the following:

1. The pseudo random number generator has desirable characteristics (e.g., a long period before repeating values).
2. The pseudo random number generator produces values that pass tests for randomness.
3. The number of repetitions of the experiment is sufficiently large to ensure accuracy of results.
4. The proper sampling technique is used (e.g., sampling with replacement vs. sampling without replacement).

When Monte Carlo methods are used in *simulation* work, the following two characteristics must be added:

1. The algorithm is valid.
2. The simulation is realistically modeled.

Applications of Monte Carlo Simulation

A Monte Carlo simulation of the probability of surviving a heart attack under various conditions was developed by Sawilowsky and Fahoome. A population of heart attack victims was identified in a certain county of a midwestern state of the United States. Information was collected on their age, where the heart attack occurred, who was present when the heart attack occurred, if the witness to the event was trained in cardiopulmonary resuscitation (CPR), if CPR was administered, if the heart presents a rhythm conducive to electric shock, decrease in survival rate per minute after incident to receiving shock, and the length of time between the attack and when the patient received electric shock treatment by medical personnel.

Probabilities for these conditions were obtained from published sources in the county where the study was undertaken. A software program was coded in Fortran 90 to conduct a Monte Carlo simulation of a heart attack victim's probability of survival based on the conditions surrounding the incident. The simulation provides estimates of survival rates when any of the above mentioned criteria are modified.

The plethora of applications using Monte Carlo simulation is very rich. A small subset of examples

from the applied literature include annealing, consumer behavior of switching brands, controlling dam water, cooling temperature of coffee, customer product ordering behavior, development of ability to perform push-ups, ecology of the Kaibab Plateau on the rim of the Grand Canyon, electromagnetism, estimating migration patterns, expected waiting times, genetic linkage, growth of yeast in a sugar solution, heroin addiction's impact on a community, image processing, inventory control, management planning, mass supply systems, material or time delays, projection of discovery of natural gas reserves, quality and reliability of products, queuing sale and consumption of commodities, systems at a 2-minute car wash, short-term forecasting, and urban growth. The list is growing.

—Shlomo S. Sawilowsky

See also Conditional Probability; Probability Sampling

Further Reading

- Marsaglia, G. (2003). Random number generators. *Journal of Modern Applied Statistical Methods*, 2(1), 2–13.
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335–341.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157–175.
- Sawilowsky, S. S. (2003). You think you've got trials? *Journal of Modern Applied Statistical Methods*, 2(1), 218–225.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, 111, 353–360.
- Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation via Fortran*. Rochester Hills, MI: JMASM.
- Sobol, I. M. (1974). *The Monte Carlo method*. Translated and adapted from the second Russian edition by Robert Messer, John Stone, and Peter Fortini. Chicago: University of Chicago Press.
- Student. (1908). On the error of counting a haemacytometer. *Biometrika*, 5, 351–360.
- Student. (1908). The probable error of a mean. *Biometrika*, 6, 1–25.

MOSAIC PLOTS

Mosaic plots were introduced by Hartigan and Kleiner as a means of visualizing contingency tables. As the name suggests, mosaic plots are pieced together from smaller rectangular tiles. Each tile corresponds to one cell of a contingency table. Its area is proportional to the size of the cell, and shape and location are determined during the construction process. We will explain the construction process using data from the *Titanic* (see Table 1 for the numbers). In this data set, age, gender, and passenger information is available for all 2,201 persons on board the *MS Titanic* on her fatal maiden voyage. Additionally, survival information is recorded for each person. We start with a single variable: Figure 1 shows a bar chart of Survival. The bars are colored according to the outcome. On the

right, a spineplot is drawn. Spineplots are variations of bar charts, where the width instead of the height of each bar is proportional to the number of cases.

Additionally, the coloring shows survival within class. We can see that survival rates go down from left to right, that is, first class passengers had the highest rate of survival (about 62%), whereas third class passengers and crew members had the lowest survival rate (about 24%). The spineplot of Figure 1 is an example of a one-dimensional mosaic plot. Mosaic plots can show many variables and are limited only by practical issues, such as screen space and interpretability.

To make the mosaic of Figure 1 two-dimensional, we can include the information on age (classified as adult or child) and draw two separate mosaic plots on top of each other, as shown in Figure 2.

The top row in this mosaic plot shows survival rates of children within the classes. It becomes apparent that there were no children in the crew—obviously a structural zero. There also were no deaths in the first two passenger classes. We can now incorporate gender in the mosaic plot by placing the mosaic of Figure 2 for each gender separately side by side, as is shown in Figure 3.

Survival rates by class, age, and gender are shown. Overall, a lot more men were on board than

Table 1 *Titanic* Data Set

Sex	Survived	Child				Adult				Total
		1st	2nd	3rd	Crew	1st	2nd	3rd	Crew	
Male	No	0	0	35	0	118	154	387	670	1364
	Yes	5	11	13	0	57	14	75	192	367
Female	No	0	0	17	0	4	13	89	3	126
	Yes	1	13	14	0	140	80	76	20	344
Total		6	24	79	0	319	261	627	885	2201

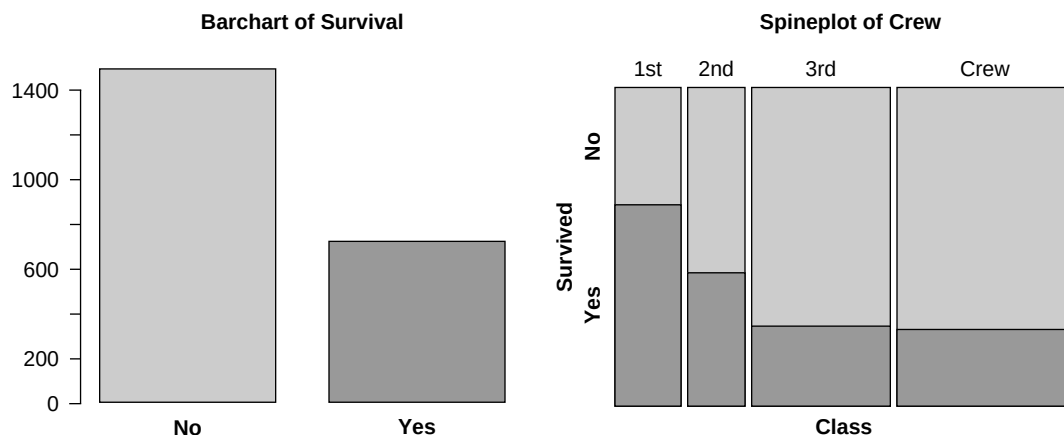


Figure 1 Bar Chart and Spineplot

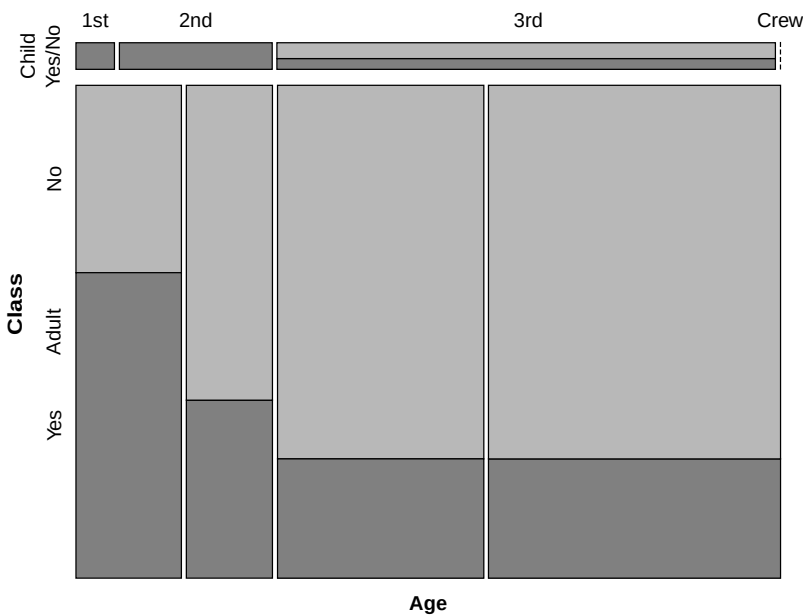


Figure 2 Mosaic Plot of Class and Age

women. Survival rates for women are significantly higher than for men. The pattern within class also looks different for women than for men. Women have a strong class/survival association: With higher class, survival improves dramatically. This is not true for

men. Survival rates of men in the second passenger class are curiously low and have led to many speculations.

The construction of mosaic plots is strictly hierarchical, which emphasizes the order of variables. Figure 4 shows the same data as the mosaic in Figure 3 with different order of the variables. The variables of class and gender are exchanged. This enables us to better compare the differences in survival rates between female and male persons on board.

Properties

Mosaic plots have excellent mathematical properties:

- The cells' sizes are visual estimates of the joint distribution of all the variables.
- Various conditional distributions are preserved.
- Links to log linear models have been established.
- Extensions of mosaics include tree maps and trellis displays.

—Heike Hofmann

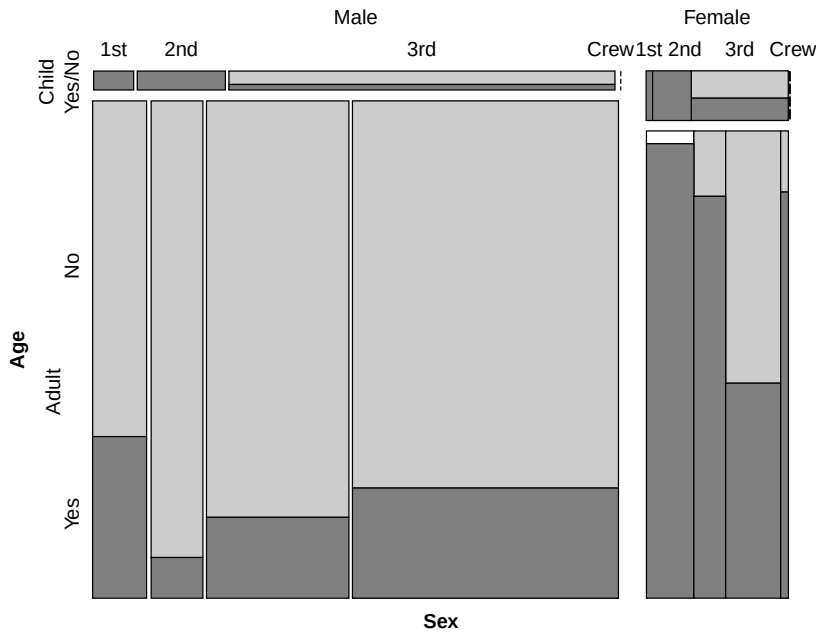


Figure 3 Mosaic Plot of Class, Age, and Gender (Sex Variable)

Further Reading

Cleveland, W. S., & McGill, M. E. (Eds.). (1988). *Dynamic graphics for statistics*. Pacific Grove, CA: Wadsworth.

Dawson, R. J. M. (1995). The “unusual episode” data revisited. *Journal of Statistics Education*, 3. Available: <http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html>

Friendly, M. (1995). Conceptual and visual models for categorical data. *American Statistician*, 49, 153–160.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8, 373–395.

Great Britain. Parliament. (1998). *Report on the loss of the S.S. Titanic: The official government enquiry*. New York: Picador.

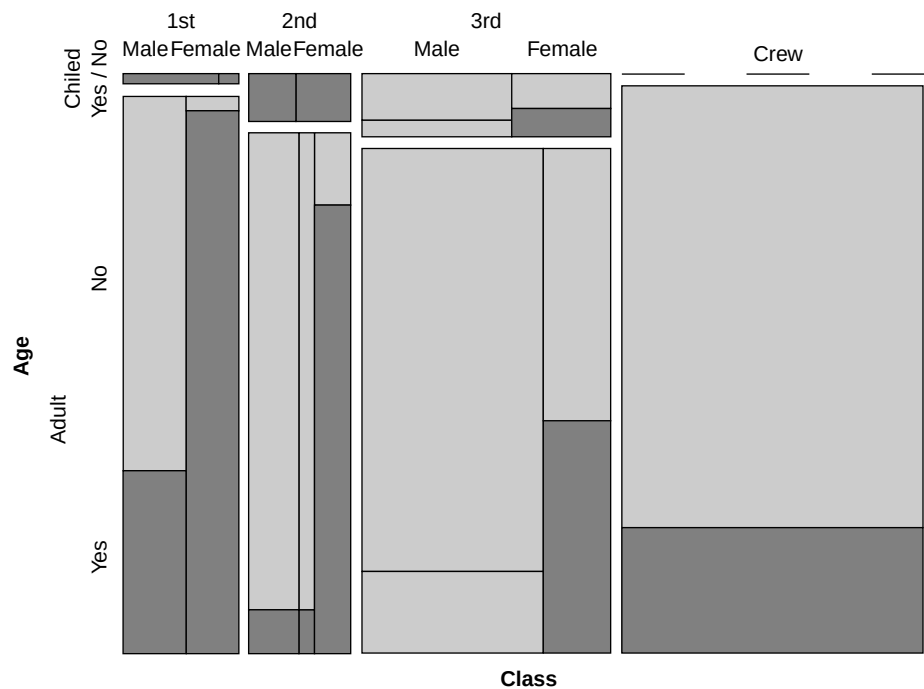


Figure 4 Mosaic Plot of Class, Age, and Gender (Class Variable)

Hartigan, J. A., & Kleiner, B. (1984). A mosaic of television ratings. *American Statistician*, 38, 32–35.

Hofmann, H. (2003). Constructing and reading mosaicplots. *Computational Statistics and Data Analysis*, 43, 565–580.

Hummel, J. (1996). Linked bar charts: Analysing categorical data graphically. *Journal of Computational Statistics*, 11, 23–33.

Manet and Mondrian data analysis tools with extensive functionality for mosaic plots in an interactive framework: <http://www.rosuda.de/software>

Mosaic plots and SAS macros: <http://www.math.yorku.ca/SCS/friendly.html#mosaics>

MOVING AVERAGE

A moving average is a simple but powerful technique that can be used on time series data or any data recorded in equal intervals, such as time. The history of moving averages goes back at least to 1924, well before computers or hand calculators were in existence. In its simplest form, it is a progressive and repetitive calculation of simple averages and does not require sophisticated computation power. However,

using a computer to do the simple calculations and to draw related graphs allows the user to focus on the interpretation of the results. The moving average has uses limited only by the imagination of the user.

It has been used as a technique to smooth data to reveal a trend and as a forecasting technique. It is part of the ARMA methodology for forecasting, where the MA stands for moving average. The use in the sophisticated ARMA forecasting process is more of a historical interpretation than an actual implementation.

The theory behind moving averages allows free interpretation and application with few details relating to theory associated with underlying statistical assumptions. It should be viewed as a descriptive statistical technique with many potential theoretical and practical uses. Because it is a very intuitive and practical technique, it is best explained by example.

Moving Average Computation and Use

The software used for illustration purposes is Minitab, a well-known, easy-to-use, and reliable software

package that runs on a PC. To start out, consider the following ordered data that were created to illustrate the calculation of a moving average (Figure 1).

Worksheet 1 ***		
↓	C1	C2
	order	data
1	1	2
2	2	3
3	3	5
4	4	7
5	5	8
6	6	9
7	7	6
8	8	5
9	9	3
10	10	2
11	11	4
12	12	6
13	13	8
14	14	10
15	15	12
16	16	8
17	17	7
18	18	5
19	19	4
20	20	3

Figure 1 Worksheet

The order refers simply to the fact that the data are in order. If the data are time series, then the order variable would be time at equally spaced intervals.

Because moving averages are often used to smooth data in the sense of eliminating nonmeaningful irregularities in the data, the best practice is to first plot the data to see what they look like. The questions that a plot can answer are, Are there any irregularities in the data that are not meaningful? Do the data display some sort of trend that might be meaningful?

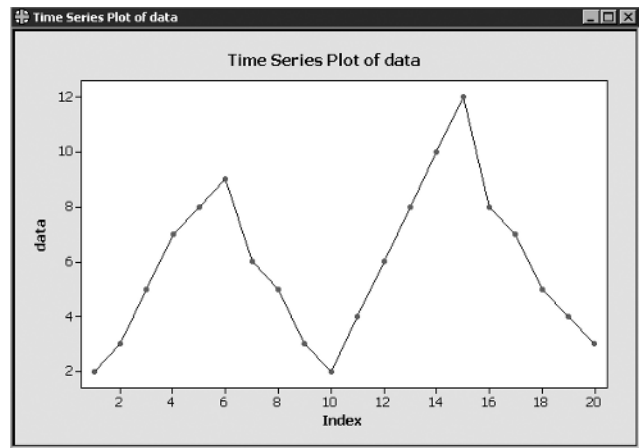


Figure 2 Time Series Plot of Data

The plot in Figure 2 was produced in Minitab by simply selecting Graph from the main menu and then Time Series Plot. Notice that the data have a slight upward trend with one valley and a couple of peaks.

To explore a moving average, consider a five-term moving average; the number of terms in the moving average is usually determined by trial and error if no other insight is available.

The hand calculations for the first term of the moving average are to simply average the first five data points:

$$\begin{aligned}
 MA(1) &= \frac{\text{data}(1) + \text{data}(2) + \text{data}(3) + \text{data}(4) + \text{data}(5)}{5} \\
 &= \frac{2 + 3 + 5 + 7 + 8}{5} = \frac{25}{5} = 5,
 \end{aligned}$$

where $MA(1)$ is the first smoothed value and $\text{data}(n)$ is the data column value at Period or Order 1.

The second smoothed value is

$$\begin{aligned}
 MA(2) &= \frac{\text{data}(2) + \text{data}(3) + \text{data}(4) + \text{data}(5) + \text{data}(6)}{5} \\
 &= \frac{3 + 5 + 7 + 8 + 9}{5} = \frac{33}{5} = 6.6.
 \end{aligned}$$

Thus, the calculations continue by shifting down one step and computing the average of the next five terms—thus the term *moving average*.

When does the calculation stop? When there are no longer five terms to average.

Once the moving average is computed, it needs to be associated at some point with the original data. There are several ways to do this depending on what the purpose of the moving average is.

One practice is to place the new moving average series of numbers one step beyond the period of the average. Thus, in the above example, MA(1) would be positioned at Order 6.

Another practice, which is, in fact, called that, is to center the moving average, so in the case of a five-term moving average, data point MA(1) would be moved to Order 3.

To see what this does in practice to the smoothing characteristics of the moving average, Minitab is used to compute both the five term moving average and the five-term centered moving average. Figure 3 shows the results of those calculations.

Worksheet 1 ***				
↓	C1	C2	C3	C4
	order	data	MA period 5	Centerd MA period 5
1	1	2	*	*
2	2	3	*	*
3	3	5	*	5.0
4	4	7	*	6.4
5	5	8	5.0	7.0
6	6	9	6.4	7.0
7	7	6	7.0	6.2
8	8	5	7.0	5.0
9	9	3	6.2	4.0
10	10	2	5.0	4.0
11	11	4	4.0	4.6
12	12	6	4.0	6.0
13	13	8	4.6	8.0
14	14	10	6.0	8.8
15	15	12	8.0	9.0
16	16	8	8.8	8.4
17	17	7	9.0	7.2
18	18	5	8.4	5.4
19	19	4	7.2	*
20	20	3	5.4	*

Figure 3 Worksheet

Note that the calculated moving averages are the same for the centered and noncentered moving averages. The only difference is with what order the series is associated. A comparison graph (Figure 4) is useful in seeing the implications of the position of the moving average.

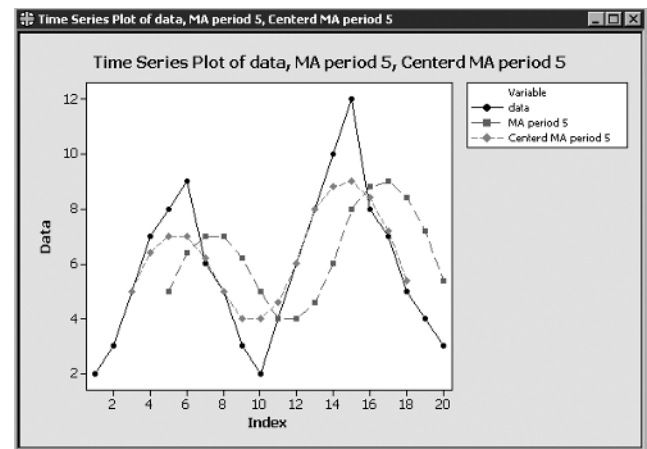


Figure 4 Time Series Plot of Data, MA Period 5, Centered MA Period 5

It is clear from the graph that the moving average has indeed smoothed the original data.

If the data represented a cycle—repeating pattern—then this smoothed version might be more representative of the true pattern. If the intent of smoothing the data is to use it to forecast a single future value, then the noncentered moving average would most likely be preferred. A naive forecasting rule, for one period beyond the actual data, is to use the last smoothed value. In this case, the forecast for Period 21, which is a future value with no data, is 5.4, the last smoothed value.

Suppose that the intent of smoothing the data is to draw out the simple upward trend. The question then is, What period of moving average is best? The answer is usually generated by trying various orders or periods (vary the number of data points averaged) of moving averages. Figure 5 shows the results of this investigation.

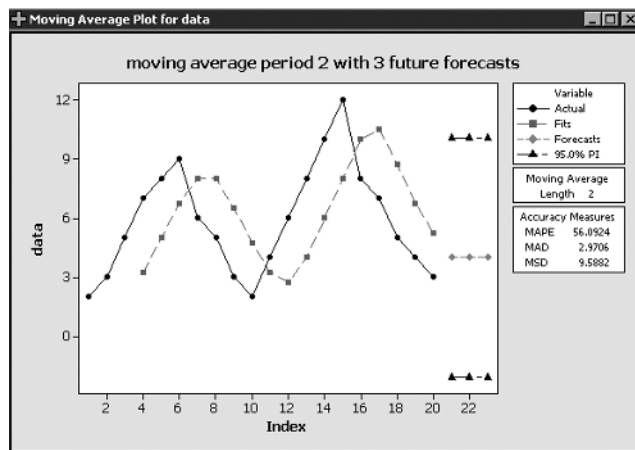


Figure 5 Moving Average Plot for Data

The first graph shows a two-period moving average with forecasts generated three periods ahead. Note that the moving average is less smooth for two terms

than with five. Also note that the forecasts generated by Minitab for three periods ahead are basically highly related to the last smoothed value. Also note that the three forecasts are identical. Thus, it can be correctly concluded that forecasts beyond one period from the original data might become less accurate. The 95% prediction interval is quite wide and indicates that the forecasts are highly variable.

Next, look at the result of using a very large number of terms in the moving average. A first guess might be that it should be smoother because it averages over many more periods. (See Figure 6.)

As might be expected, the larger period moving average results in a much smoother line, which is reflecting the trend of the data without the peaks and valleys of the original data. The forecasts for three periods ahead are still the same values for all three forecasts; a moving average is not considered useful for more than short term forecasts.

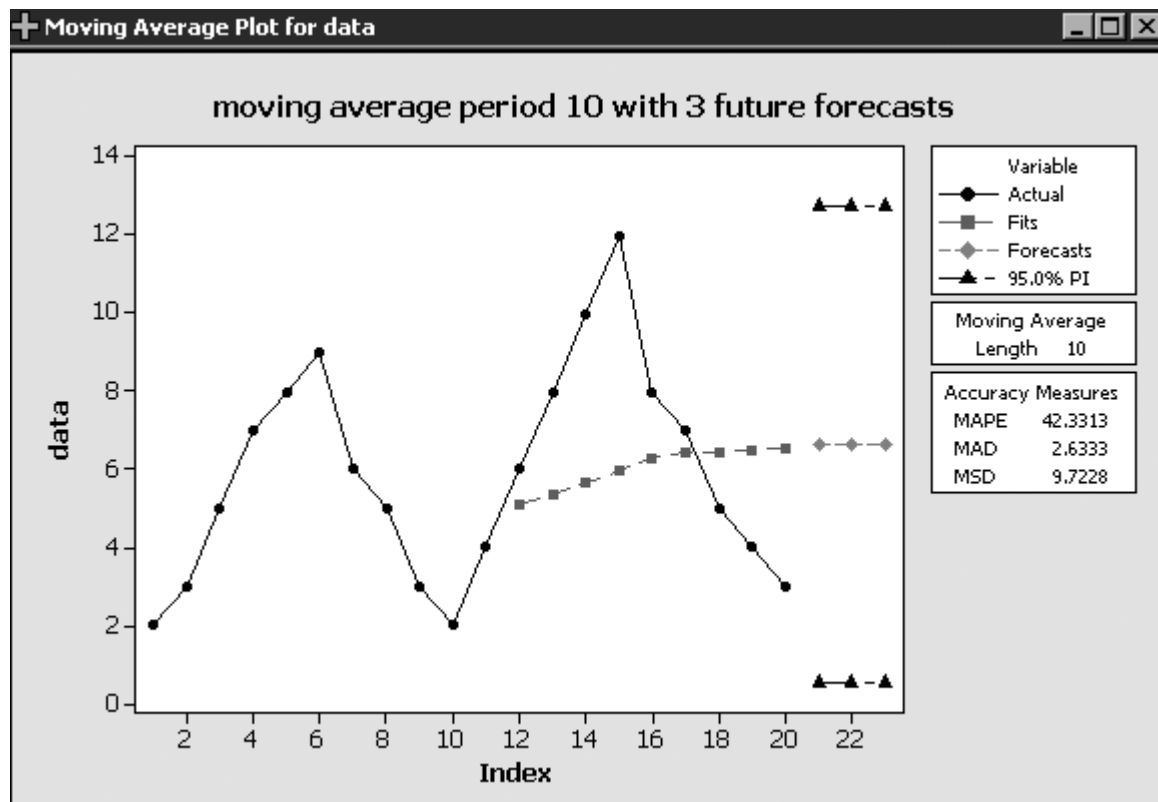


Figure 6 Moving Average Plot for Data

If the goal of using a moving average is to simply smooth the data to pull out the trend, then that has certainly happened with these data. The moving average line is seen to be a very linear trend that is increasing slightly.

To get an overall look at the result of adding more terms to the moving average, Figure 7 is useful.

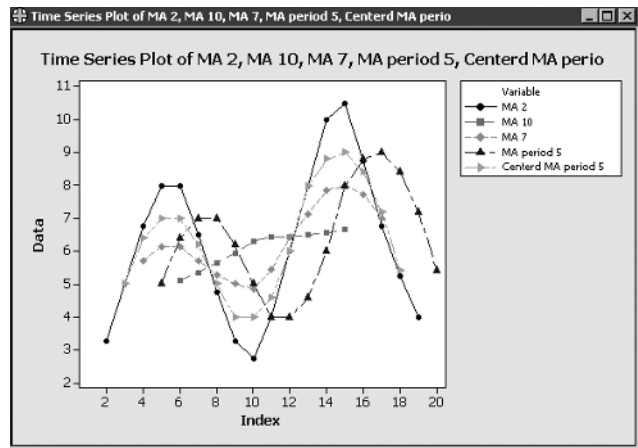


Figure 7 Time Series Plot of MA 2, MA 10, MA 7, MA Period 5, Centered MA Period

The trend is obviously most apparent when many terms, 10, are averaged. The centered moving average is very good at simply smoothing the existing series and representing the smoothed data consistent with the original data—that is, not shifted. What is correct? Well, there are no incorrect or absolutely correct results. It depends on the use of the moving average. If the goal is to smooth the data to help interpretation or to fit future models, then centering the data is best. If the goal is to produce a general trend in the data, then a higher order number of terms is most likely the best solution.

Applications

Aside from a simple technique to smooth the data or to smooth and produce a simple trend, what are moving averages good for?

There is a classical result that has proved by usage to be very profound for moving averages. The typical application has been in the stock market, where it is desirable to predict changes over time in the upward or downward direction. Although complex statistical models have some use in predicting instantaneous changes (turning points) in stock market performance, few simple techniques can give advice on price movements. Moving averages have been used to give stock performance information with relatively little technical analysis or input.

Consider the following moving average that represents a 50- and 100-period moving average on stock price (Figure 8).

When the stock price falls below the 50- or 100-day moving average, it is recommended that the stock owner sell. Note that if the owner had sold at the point at which the stock went below the moving average, a long period of low prices would have been avoided.

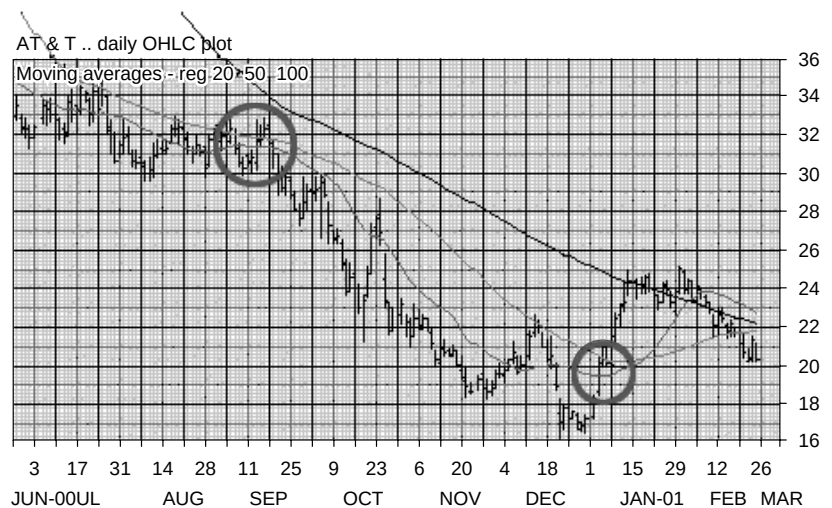


Figure 8 Moving Average Plot for a 50 and 100 Period

Source: <http://www.investopedia.com/university/movingaverage/default.asp>

When the stock goes above the 50- or 100-period moving average, the stock may be purchased with the hope of a long upward trend, as shown on the graph.

How does the individual use this information? The investor can update and plot the moving average, usually over a 49 or longer (days or months as appropriate) period of time. This long period will give a general trend for the data, as seen in the previous example. It is easy to update a long-term moving average one day at a time because one has to remove only the first number and add the last new period to the total and divide by the number averaged. Dedicated investors have used this type of updating along with hand plotting of the averages for many years without calculators or computers. The rule of thumb is that when the stock price drops below the long-term moving average, sell, and when it goes above the long-term moving average, buy. Although this will hardly ever hit the exact turning points in stock prices, over the long term, it has proved to be a very good decision rule that is easy to use. It works especially well with mutual funds, which are a mixture of stocks.

How would these results apply to other ordered data? It might prove very useful in evaluating education systems. Some states require all students in grade school to take end-of-year performance tests. Over a period of time these tests could be recorded and smoothed. The moving average could be used to show trends in the data: A long-term moving average could be used to indicate potential trouble when scores fall below the moving average, or improvement when scores rise above a moving average.

Another example might be in monitoring student weight gain or loss as related to the removal or addition of soft drink dispensing machines in public schools. Suppose student average weight is recorded for several months and years for a particular school. Then suppose at some point in time, soft drink machines were introduced in the school. If the average student weight went above the long-term moving average around the point that the soft drink machines were introduced, that would be an indication that they might be contributing to weight gain among students. If soft drink machines are removed from schools and the average weight goes below the long-term moving

average, that would supply evidence that the program was working.

Conclusions

A moving average is a simple statistical method that can prove very useful in monitoring and predicting data that are recorded in an ordered or time based manner. The basic calculations can be done easily by hand but are also aided by computer calculation. It is a technique that can deal with time or ordered data in an intuitive manner and should be part of the data analysis toolbox of anyone involved in working with data. It has few theoretical restrictions, and its application is limited only by the ingenuity of the user. Several statistical software packages that include spreadsheets can compute moving averages, but if a computer isn't available, they can be computed by hand.

—Marietta J. Tretter

See also Mean; Measures of Central Tendency

Further Reading

- Hanke, J. E., & Wichern, D. W. (2005). *Business forecasting*. Upper Saddle River, NJ: Pearson.
- Kendall, M. G., & Stuart, A. (1966). *The advanced theory of statistics*, Vol. 3. New York: Hafner.

Moving averages: <http://www.investopedia.com/university/movingaverage/default.asp>

MULTICOLLINEARITY

Multicollinearity is a phenomenon that may occur in multiple regression analysis when one or more of the independent variables are related to each other. The relationship between the independent variables could be expressed as near linear dependencies.

A simple form of multicollinearity could be due to high correlation between some pair(s) of independent variables. Assuming that there are five independent variables, X_1 , X_2 , X_3 , X_4 , and X_5 , that are being used to

develop a model for the dependent variable, Y , suppose that the correlations between X_2 and X_3 , and between X_4 and X_5 , are high (≥ 0.90). In such a case, maybe one of the two variables X_2 or X_3 , and one of the two variables X_4 or X_5 , could be used in the model. The rationale is that there is very little independent contribution of the variable that is left out, in the presence of the one that is retained, in explaining the variation in the dependent variable.

More complex forms of multicollinearity may exist when three or more independent variables are nearly linearly related, even though there might not be a high pair wise correlation among them. This could be represented by the approximate relation $X_1 + 2X_2 - X_3 \simeq 0$.

Effects of Multicollinearity

The impact of multicollinearity on the development of a multiple regression model and drawing inferences from it is multifaceted. First, the variance of the estimated model coefficients is large, which leads to their instability. Small perturbations of the observations and/or omission of an independent variable from a large set of independent variables may cause large fluctuations in the estimated regression coefficients.

Second, the signs associated with the regression coefficients may be somewhat contrary to what is expected, given the setting of the problem. For example, in predicting resource requirements (Y), using production quantity (X_1), direct labor (X_2), and raw material (X_3), it is normally expected that the coefficients associated with X_1 , X_2 , and X_3 will be positive. However, given that X_1 and X_2 are highly correlated, and so are X_1 and X_3 , it is possible for one of the estimated coefficients to turn out negative.

Third, it is possible for the full model to be statistically significant, even though none of the individual coefficients is significant. Thus, all of the independent variables, taken collectively, may provide a good fit to the response variable, leading to a small value of the residual sum of squares. On the contrary, individual coefficients are estimated poorly. In the presence of multicollinearity, a given regression coefficient may not reflect the inherent

effect of the particular regressor. It is influenced by the variables that are in the model.

Detection of Multicollinearity

All of the three possible effects, previously discussed, may be examined closely for the possible presence of multicollinearity. These are as follows: high pairwise correlation among some of the independent variables; wide confidence intervals associated with the regression coefficients; opposite sign of estimated regression coefficients based on theoretical knowledge of the problem; nonsignificant tests on individual coefficients while the model as a whole is significant; and large changes in the estimated coefficients when an observation is slightly changed or deleted, or when an independent variable is added or deleted.

Another method of detecting multicollinearity is through computation of the variance inflation factors (VIFs). A VIF measures the degree to which an independent variable is related to the other independent variables. Consider regressing the independent variable (X_k) to the other independent variables ($X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_p$). The variance inflation factor for X_k is given by

$$(\text{VIF})_k = \frac{1}{1 - R_k^2}, \quad k = 1, 2, \dots, p,$$

where

p represents the number of independent variables and

R_k^2 is the coefficient of multiple determination when the variable X_k is regressed on the other remaining independent variables.

It is observed that when $R_k^2 = 0$, implying that X_k is not linearly related to the other independent variables, $(\text{VIF})_k$ is equal to 1. As R_k^2 becomes large and closer to 1, $(\text{VIF})_k$ becomes large. In the extreme case when $R_k^2 = 1$ (i.e., X_k is perfectly linearly related to the other independent variables), $(\text{VIF})_k$ is unbounded. The largest $(\text{VIF})_k$, among all the regressors, is frequently used as an indicator of the degree of multicollinearity. As a rule of thumb, a maximum $(\text{VIF})_k$ of 10 or

more is considered an indication of the presence of multicollinearity.

Remedial Measures for Multicollinearity

On examination of the correlation matrix of the independent variables, one or more independent variables, which are highly correlated to others, may be dropped from the model. This measure has some drawbacks. First, if data have been collected on all the regressor variables, by dropping one or more of these, no direct information is obtained about the dropped variables. Furthermore, the magnitude of the regression coefficients retained in the model is affected by the correlated regressors not included in the model.

In polynomial models, expressing the independent variable in the form of deviation from the mean may reduce multicollinearity among the first, second, or higher order terms.

Another approach could be to use a biased estimation process compared to least squares regression. In this context, ridge regression is an alternative. Although it introduces some bias in the estimator, this procedure has a much smaller variance. The net impact, as measured by the mean squared error, which is the expected value of the squared difference between the estimator and the true parameter value, could be smaller. The mean squared error can be expressed as the sum of the variance and the squared bias of the estimator. For an unbiased estimator, such as the least squares estimator, the mean squared error equals the variance of the estimator.

—Amitava Mitra

See also Assessment of Interactions in Multiple Regression; Curvilinear Regression; Partial Least Square Regression; Regression Analysis

Further Reading

Myers, R. A. (2000). *Classical and modern regression with applications*. Boston: Duxbury.

Multi-collinearity—Variance inflation and orthogonalization in regression: <http://creative-wisdom.com/computer/sas/collinear.html>

MULTIDIMENSIONAL APTITUDE BATTERY

The Multidimensional Aptitude Battery (published by Sigma Assessment Systems) was developed by Douglas N. Jackson and is currently in its second edition (MAB-II). Following the tradition established by Robert Yerkes during World War I with the Army Alpha and Beta for testing literate and illiterate recruits, respectively, the MAB categorizes abilities into verbal and nonverbal components. The MAB contains five verbal subtests (Information, Comprehension, Arithmetic, Similarities, and Vocabulary) and five nonverbal subtests (Digit Symbol, Picture Completion, Spatial, Picture Arrangement, and Object Assembly). Each of these 10 subtests is in a five-point multiple-choice format and has a 7-minute time limit. With instructions and practice items, the MAB-II takes about 90 minutes to complete. With the exception of Digit Span, which Wechsler adapted from the Binet, the MAB contains the same subtests as included in the Wechsler Bellevue, Wechsler Adult Intelligence Scale (WAIS), and Wechsler Adult Intelligence Scale Revised (WAIS-R). Correlations between corresponding subtests of the MAB and WAIS-R are of the same order of magnitude as the correlations between subtests of the WAIS and WAIS-R.

Subtest raw scores are converted to scaled scores with a *T* score metric (i.e., mean of 50, standard deviation of 10). The five verbal and five nonverbal scale scores are summed to produce a sum of scale scores for Verbal and Performance IQ, respectively. The sum of all 10 scale scores is used to calculate Full Scale IQ. Age groups for obtaining the three IQ scores are categorized into 16–17, 18–19, 20–24, 25–34, 35–44, 45–54, 55–64, 65–69, and 70–74. These nine age groups are also used to obtain age corrected scale scores for the 10 subtests. The three IQ scores have a mean of 100 and a standard deviation of 15.

The MAB-II can be administered in paper and pencil format with reusable question booklets and answer sheets to individuals or to groups. Where large volumes of clients are assessed comprehensively (i.e., 10 subtests) in contexts such as employee selection or

research, group administration makes such testing feasible. There are separate question booklets for the verbal subtests and performance subtests. In addition to hand scoring and mail in scoring, computerized administration with automated scoring and reporting is also available. Reports generated by the software are either clinical with some limited interpretation or ASCII (text) data file reports for research. The MAB-II is available in French and English.

—John R. Reddon

See also Ability Tests; Aptitude Tests; Intelligence Quotient; Intelligence Tests

Further Reading

- Bracy, O. L., Oakes, A. L., Cooper, R. S., Watkins, D., Watkins, M., Brown, D. E., et al. (1999). The effects of cognitive rehabilitation therapy techniques for enhancing the cognitive/intellectual functioning of seventh and eighth grade children. *International Journal of Cognitive Technology*, 4, 19–27.
- Vernon, P. A. (2000). Recent studies of intelligence and personality using Jackson's Multidimensional Aptitude Battery and Personality Research Form. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 195–212). Norwell, MA: Kluwer.

Multidimensional Aptitude Battery-II: <http://www.sigmaassessmentssystems.com/assessments/mab.asp>

MULTIPLE AFFECT ADJECTIVE CHECKLIST–REVISED

The Multiple Affect Adjective Checklist–Revised (MAACL-R, published by EdiTS/Educational and Industrial Testing Service) is a trait state test for the measurement of negative and positive affects. Both forms consist of 132 adjectives, but the instructions for the state form are to “check all adjectives which describe how you feel ‘now’ or ‘today,’” whereas those for the trait form say “check all the adjectives which describe how you ‘generally feel.’” Scales for both forms are provided for anxiety (A), depression

(D), hostility (H), dysphoria (Dys = the total A + D + H), sensation seeking (or surgency) affect (SS), and total positive affect (PASS = PA + SS). The negative affect scales were originally developed using an empirical item selection method, but the scales in the revised version are a result of factor analyses of both state and trait forms. In order to control acquiescence response set (total number of items checked), *T* scale norms are provided for each of three ranges of numbers checked. Norms are available for general adult, adolescent, college student, and elderly populations. The general adult norms are based on a national representative sample. There are also scales for random responding and intentional response manipulation (faking). The median reading level for the 132 adjectives is sixth grade, and 90% of the adjectives in the scales are at or below the eighth grade level.

Reliability

Internal (alpha) reliabilities have been determined within a wide variety of populations for both the state and trait forms. These reliabilities are high; most are .80 or higher. The highest reliability is for the Dys scale, which is .90 or higher in nearly all of the samples for the state form and all for the trait form. In conformance with the trait state distinction, retest reliabilities are expected to be high for the trait form but not for the state form. The state form of the PA scale is the only one showing low but consistently significant retest reliability. In contrast, retest reliabilities for the trait are moderately high even after an 8 week interval. Correlations between corresponding state and trait scales on a single day are significant but low, with the exception of the PA scale, which is high. Aggregated or averaged state scores over a few days to a week correlate more highly with trait scores.

Validity

The state scales have been used to measure change in many situations, including examination induced anxiety, aggression provocation, group therapy, listening to music, exercise, videos, and Air Force basic training, where it predicts discharge status after training.

The trait form correlates with many other scales for emotions, and peer and observer ratings. In a clinical diagnostic study, a combination of scales differentiated normals, schizophrenics, depressives, and patients with anxiety disorders.

—Marvin Zuckerman

Further Reading

Lubin, B., Zuckerman, M., Hanson, P. G., Armstrong, T., Rinck, C. M., & Seever, M. (1986). Reliability and validity of the Multiple Affect Adjective Check List-Revised. *Journal of Psychopathology and Behavioral Assessment*, 8(2), 103–117.

EdiTS/Educational and Industrial Testing Service: <http://www.edits.net>

MULTIPLE-CHOICE ITEMS

The multiple-choice (MC) item is the most common format for questions on educational tests. MC items are also used in other domains, including personality, interest, and attitude inventories and surveys. There are several forms of MC items and many guidelines available to item writers. Thomas Haladyna has provided leadership in this area, promoting a systematic approach to the development and validation of MC items.

In the design of any measurement instrument, two primary issues must be addressed early: what to measure and how to measure it. Both questions inform the choice of item format. Clearly, the construct being measured and the nature of the cognitive task required to measure the construct will play a role in deciding whether the MC format is appropriate.

The strengths of MC items were first debated in the literature shortly after the introduction of the Army Alpha test, a recruit screening and classification instrument used during World War I. Previously, aptitude and intelligence related measurement procedures included interactive tests administered one on one. It was proposed that test items could be constructed to be self-administered and yield sufficient

information about knowledge, skill, or level of ability or some other trait. The MC item allowed test developers to cover a wide range of content and skills and design instruments that could be group administered. With advances in machine scoring during the 1950s, the MC item quickly became the predominant testing format.

Measurement specialists now caution test designers and argue that validity, not convenience, should determine what types of item formats should be employed on a given measurement instrument. The objective is to craft the item in order to obtain the best measure of the construct or cognitive task of interest. The MC item offers a great deal to test designers because of its flexibility. However, as with any measurement procedure, MC items can be poorly written, leading to confusion and misinterpretation by the test taker, and thus scores of limited reliability and validity.

Forms of MC Items

The typical MC item consists of a stem and three or more options, where one option is the correct option and the others are incorrect options or distractors. The stem is the question or opening statement that is answered or completed by the options. Consider the following item:

1. What ratio of variances describes the reliability of test scores?
 - A. Errors of measurement and observed scores
 - B. Errors of measurement and true scores
 - C. True scores and observed scores

The question “What ratio of variances . . .” is the stem. The three choices that follow (A, B, and C) are the options. Option C is the correct option, and Options A and B are distractors or incorrect options. It is possible to construct this item as a completion statement:

2. The reliability of scores is regarded as the ratio between the variances of
 - A. errors of measurement and observed scores.
 - B. errors of measurement and true scores.
 - C. true scores and observed scores.

Both forms of the typical MC item have gained support from the measurement community, and neither is viewed as yielding greater validity than the other. However, there is some preference for the complete question format because the stem is more likely to contain the main idea of the item (an important guideline for MC items).

A slight variant of the typical MC item is the alternate choice item. This is identical to the typical MC item except that there are only two options. Another variant is the multiple true false, where there may be many options and each is evaluated in terms of being true or false. Another MC format is the complex MC or Type K format. In this format, there is a stem followed by potential answers, followed by a set of options to which the test taker responds. Consider this example:

3. When considering the need to design a test that will cover a lot of content in a brief period of time, which of the following item types could be employed effectively?
 1. Completion items
 2. Essay items
 3. Multiple-choice items
 4. True-false items
- A. 1 and 2
B. 1, 3, and 4
C. 3 and 4

In this example, the options (A, B, and C) are composed of combinations of the potential answers to the question (1 to 4). This MC item format tends to be more difficult and is commonly found in the health sciences, where the object is to differentiate the appropriateness of various combinations of options (for example, combinations of health related symptoms).

On the question of optimal number of options, there is substantial evidence suggesting that for most purposes and in most settings, three options are optimal. The evidence suggests that three option items provide no greater chance of correct guessing than items with more options because, more so than not, additional options are less plausible and are not considered by most students, and that few, if any, individuals actually engage in blind guessing. Empirical

research has provided evidence that few four- or five-option items have that many functional distractors, essentially making the item a three-option or, in many cases, a two-option item. In addition, item writing effort and time can be reduced and more three-option MC items can be asked, allowing the test designer to include more items and improve content coverage.

MC Item Writing Guidelines

Guidelines are available for the MC item writer that consist of issues related to the stem, the options, and other general editorial and content concerns. These guidelines are covered to various extents by measurement textbook authors. Haladyna provides the most thorough review of guidelines available. Some of the guidelines include concerns about clarity of word choice and writing style. Content concerns include the idea that each item should regard a single idea on an important topic that is not opinion based (unless the focus of the item regards a specific opinion). Guidelines regarding the stem focus on the stem containing the main idea of the item without excessive wordiness and the selective use of formatting to bring attention to words that might be missed, such as NOT.

Most of the available guidelines cover issues related to the options. This points to the relative importance that most measurement specialists place on the role of the options in the quality of the item. Options should be independent, similar in content and grammatical structure, equal in length, and free of clues to the correct option (e.g., options that resemble or contain words also present in the stem). In most cases, options such as “none of the above” and “all of the above” should be avoided, as should negative options or options containing words such as “not.” Perhaps most importantly, only one option should be correct, and distractors (incorrect options) should be plausible. One of the best ways to achieve this is to use typical errors or misconceptions as distractors. It is also recognized that one way to control item difficulty is to vary the proximity or plausibility of the distractors. The closer in similarity the options, the more difficult the item.

Advantages and Disadvantages of MC Items

MC items have several advantages and disadvantages compared to constructed response item formats. Among the advantages, MC items can provide direct assessment of a wide range of skills, including higher order thinking and complex cognitive tasks. To achieve these goals, item writers need training and experience. MC items are efficient to administer and objectively scored, and they provide a mechanism for broad sampling of the content domain or the construct. Analysis of distractors selected by test takers provides a way to obtain diagnostic information. When distractors are designed to include common misconceptions or errors, knowing which distractors are being selected provides information on which misconceptions are still held or where additional instruction may be helpful.

Among the disadvantages, MC items provide limited or indirect measurement of some skills, including the abilities to recall, provide an explanation or examples, express or organize ideas, generate a novel solution, or construct something. Other formats are better suited for these skills. Because the number of MC options is typically small, the content of any one item tends to be based on artificially structured knowledge and is rather closed to interpretation; however, we recognize that a series of well constructed MC items may approach these objectives effectively. When reading is not the target objective, reading skills may interfere with responses to MC items and the possibility of correctly guessing is always present (but may be minimized with the use of effective distractors).

—Michael C. Rodriguez

See also Completion Items; Essay Items; True/False Items

Further Reading

- Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple choice item writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334.

Rodriguez, M. C. (2005). Three options are optimal for multiple choice items: A meta analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.

MULTIPLE COMPARISONS

In data analysis, the question of interest rarely can be answered by a single statistical test or comparison. The term *multiple comparisons* describes an analysis that involves more than one statistical test or comparison based on the same set of data.

Why Multiple Comparisons Need Special Treatment

Multiple comparisons add an additional level of complexity to the analysis. For any individual comparison or test, you set an alpha level that determines the probability of a Type I error (i.e., concluding that the difference is significant when, in reality, it is not).

Suppose you choose a customary alpha of .05 (5%). Now, if you perform two independent comparisons using the same alpha, the chance of making a Type I error on each of them is 5%, and the chance of avoiding a Type I error on each is 95%. So, the chance of avoiding a Type I error altogether is $0.95 \times 0.95 = 0.9025 = 90.25\%$. The complement of this, 9.75%, represents the probability of making at least one Type I error somewhere in the set of comparisons. This probability is sometimes called the *familywise alpha level* (or the *familywise error rate*).

You can see that the familywise alpha is significantly higher than the nominal alpha we specified as 5%. The situation gets worse as the number of comparisons increases: With five comparisons, the familywise alpha is about 22.6%. The general formula for calculating the familywise alpha is

$$\alpha_{\text{familywise}} = 1 - (1 - \alpha_{\text{nominal}})^k,$$

where k is the number of comparisons.

Figure 1 shows how familywise alpha varies in relation to the number of comparisons. By the time

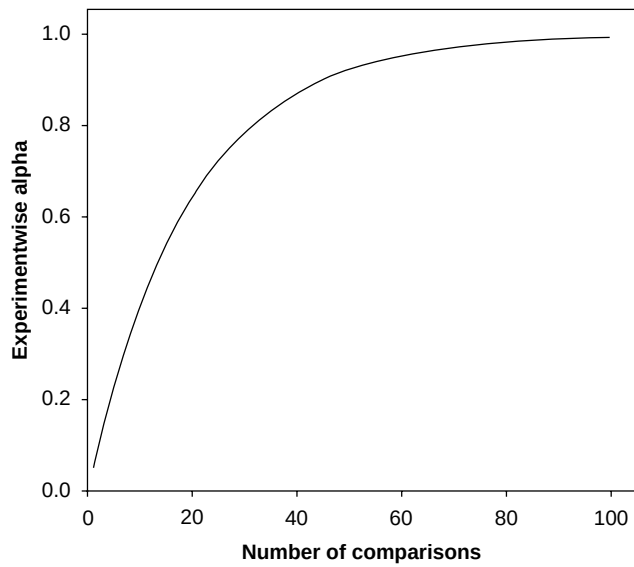


Figure 1 Experimentwise Alpha Based on Number of Comparisons

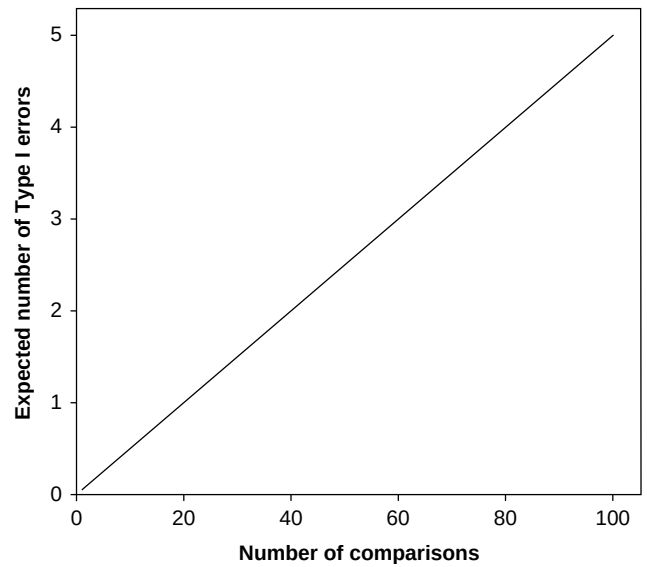


Figure 2 Expected Number of Type I Errors Based on Number of Comparisons

you perform about 50 comparisons, you are almost certain to commit at least one Type I error. Fifty comparisons may sound like a lot, but consider that if you scan a correlation matrix for 10 variables for significant correlations, you are performing 45 tests.

Figure 2 shows the expected number of Type I errors based on the number of comparisons. The more comparisons, the more Type I errors. In the hypothetical example of a correlation matrix for 10 variables, you should expect to find two to three significant

Correlations

Pearson Correlation

	random 1	random 2	random 3	random 4	random 5	random 6	random 7	random 8	random 9
random 2	-.147								
random 3	.145	-.035							
random 4	-.049	.019	-.038						
random 5	-.185	.131	-.117	.007					
random 6	.039	.007	-.023	.100	-.052				
random 7	.120	-.093	-.008	-.067	-.079	-.017			
random 8	-.117	-.178	.030	-.030	-.231*	-.093	-.073		
random 9	.041	.128	-.027	.063	.136	-.150	.046	-.125	
random 10	-.240*	.069	-.064	-.079	-.084	.009	.000	.087	.085

* Correlation is significant at the 0.05 level (2-tailed).

Figure 3 Correlation Matrix of 10 Randomly Generated Variables, Showing Two Type I Errors (Variable Pairs (5,8) and (1,10))

(but spurious) correlations by chance alone, even if all the variables are independent.

Approaches for Handling Multiple Comparisons

To avoid the problem of familywise alpha exploding as the number of comparisons increases, various methods have been developed to control the familywise alpha. Most methods center around selecting a conservative nominal alpha level, so that the familywise alpha is controlled given the number of comparisons to be made.

In controlling the familywise alpha, it is important to account for the right number of comparisons. In some cases, this will *not* be the same as the actual number of tests or comparisons performed, as explained below.

Planned Versus Post Hoc Comparisons

Comparisons or tests in a statistical analysis can be broken down into two categories, *planned* and *post hoc*, based on how and when the decision is made to perform the specified comparisons or tests.

Planned comparisons are comparisons that are specified before the data are collected or analyzed, based on the theoretical foundations of the research (or the results of previous studies). A typical example would be a clinical study where each of several treatment groups is compared to a baseline or control group. In this case, there are $\frac{k!}{2(k-2)!}$ possible comparisons, but you are only performing $(k-1)$ of them, where k is the number of groups including the control group. Because you are restricting the number of comparisons, you needn't be as conservative as you would if you were performing all possible comparisons. In fact, some researchers argue that as long as the number of planned comparisons is small, no special handling of multiple comparisons is needed.

Post hoc comparisons are comparisons that are made after the data have been analyzed and examined for patterns. To extend the example above, suppose we have five treatments and a control. After performing the planned comparisons as described above, we find that two of the five treatments show significant improvement over the baseline condition. If we want to test whether those two specific treatments are significantly different from each other, this is considered

a post hoc comparison because we didn't decide to make the comparison until after we noticed that those two treatments were significantly different from the control. In the case of post hoc comparisons, you should be more conservative than with planned comparisons. Because you didn't know what comparisons you were going to make up front, any of the possible comparisons might have looked interesting depending on the data, so you need to plan as if all possible comparisons were under consideration. In this case, you'd need a method that considered all $\frac{k!}{2(k-2)!}$ potential comparisons in adjusting the familywise alpha level, even though you're making only one post hoc comparison.

Familywise Error Rate and Power

As in the case of individual hypothesis tests, in a multiple comparison situation there is a trade off between Type I error rate and power. As you reduce your Type I error rate (alpha), you also reduce your power, or ability to detect true differences. This is especially true for multiple comparison procedures, where the nominal alpha often must be quite small to control the familywise alpha. Most multiple comparison procedures attempt to maximize power by using dependence among tests and other characteristics of the research design, while still controlling the familywise error rate.

Specific Procedures

Several popular procedures for performing multiple comparisons are briefly described below. Some of these methods have their own entries in the encyclopedia.

Bonferroni

A very conservative approach that sets the per comparison alpha to $\alpha_{\text{familywise}}/k$, where k is the number of comparisons. This is also known as the Dunn test or Dunn correction.

Fisher's Least Significant Difference

Fisher's least significant difference test takes a different approach to comparing multiple group means.

It is a two stage test. First, the overall analysis of variance (ANOVA) test is performed. If the overall test (the omnibus test) is not significant, then no specific mean comparisons are made. If the omnibus test is significant, then the means are compared using unadjusted test values. The significance of the omnibus test provides some protection against Type I errors by ensuring that at least some of the mean differences are definitely significant.

Tukey's Honestly Significant Difference

This test involves calculating the smallest difference that would be significant, controlling for familywise alpha. The formula is

$$d_{\text{Tukey}} = \frac{q_T \sqrt{\text{MSE}}}{\sqrt{n}},$$

where q_T is the studentized range statistic based on the desired familywise alpha, the number of groups (*not* the number of comparisons), and the sample size; MSE is the mean squared error for the overall ANOVA; and n is the sample size for each group. Any mean difference that is larger than this value is deemed significant.

Tukey-Kramer

This is an adaptation of the Tukey test to handle comparisons where the groups being compared have unequal sample sizes. In this case, the reciprocal of the harmonic mean of the individual group sample sizes is used in place of the equal sample size n :

$$n' = \frac{k}{\left(\frac{1}{n_1} + \frac{1}{n_2} + \dots + \frac{1}{n_k}\right)}$$

The test value d_{Tukey} is calculated as described above using n' instead of n , and testing proceeds as usual.

Scheffé

This procedure controls the familywise error rate by adjusting the critical F value used to perform

the post hoc tests. The test F statistic is calculated in the usual way for each comparison, and the critical value is calculated by multiplying the critical value for the overall test of differences among all the means (the omnibus test) by $(k - 1)$, where k is the number of groups (*not* the number of comparisons).

Dunnett

This test applies to planned comparisons in the specific case where the mean for each of a set of treatment groups is being compared to the mean for a control group. This test is similar to the Tukey HSD test and involves calculating the smallest difference that would be significant, controlling for familywise alpha, assuming the specific set of control treatment comparisons. The formula is

$$d_{\text{Dunnett}} = \frac{q_D \sqrt{2 \cdot \text{MSE}}}{\sqrt{n}},$$

where q_D is the critical value for Dunnett's test based on the desired familywise alpha, the number of groups (*not* the number of comparisons), and the sample size; MSE is the mean squared error for the overall ANOVA; and n is the sample size for each group. Any mean difference between a treatment group and the control group that is larger than this value is deemed significant.

—Clay Helberg

See also Bonferroni Test; Dunn's Multiple Comparison Test; Fisher's LSD; Tukey-Kramer Procedure

Further Reading

- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.
- Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. New York: Chapman & Hall.
- Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, 96, 589–596.
- NIST/SEMATECH. (n.d.). How can we make multiple comparisons? *NIST/SEMATECH e-Handbook of statistical methods* (§7.4.7). Retrieved from <http://www.itl.nist.gov/div898/handbook/prc/section4/prc47.htm>

Toothaker, L. E. (1993). *Multiple comparison procedures* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-089). Newbury Park, CA: Sage.

Multiple Comparison Procedures outline, by Gerald Dallal:
<http://www.tufts.edu/~gdallal/mc.htm>

MULTIPLE CORRELATION COEFFICIENT

The multiple correlation coefficient generalizes the standard coefficient of correlation. It is used in multiple regression analysis to assess the quality of the prediction of the dependent variable. It corresponds to the squared correlation between the predicted and the actual values of the dependent variable. It can also be interpreted as the proportion of the variance of the dependent variable explained by the independent variables. When the independent variables (used for predicting the dependent variable) are pairwise orthogonal, the multiple correlation coefficient is equal to the sum of the squared coefficients of correlation between each independent variable and the dependent variable. This relation does not hold when the independent variables are not orthogonal. The significance of a multiple coefficient of correlation can be assessed with an F ratio. The magnitude of the multiple coefficient of correlation tends to overestimate the magnitude of the population correlation, but it is possible to correct for this overestimation. Strictly speaking, we should refer to this coefficient as the *squared* multiple correlation coefficient, but current usage seems to ignore the adjective “squared,” probably because mostly its squared value is considered.

Multiple Regression Framework

In linear multiple regression analysis, the goal is to predict, knowing the measurements collected on N subjects, a dependent variable Y from a set of J independent variables denoted

$$\{X_1, \dots, X_j, \dots, X_J\}. \quad (1)$$

We denote by $\mathbf{X}$ the $N \times (J + 1)$ augmented matrix collecting the data for the independent variables (this

matrix is called augmented because the first column is composed only of ones), and by $\mathbf{y}$ the $N \times 1$ vector of observations for the dependent variable. These two matrices have the following structure:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,J} \end{bmatrix}$$

and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix}.$ (2)

The predicted values of the dependent variable $\hat{Y}$ are collected in a vector denoted $\hat{\mathbf{y}}$ and are obtained as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad \text{with} \quad \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3)$$

The regression sum of squares is obtained as

$$SS_{\text{regression}} = \mathbf{b}^T\mathbf{X}^T\mathbf{y} - \frac{1}{N}(\mathbf{1}^T\mathbf{y})^2 \quad (4)$$

(with $\mathbf{1}^T$ being a row vector of 1s conformable with $\mathbf{y}$).

The total sum of squares is obtained as

$$SS_{\text{total}} = \mathbf{y}^T\mathbf{y} - \frac{1}{N}(\mathbf{1}^T\mathbf{y})^2. \quad (5)$$

The residual (or error) sum of squares is obtained as

$$SS_{\text{error}} = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}. \quad (6)$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation, denoted $R_{Y,1,\dots,J}^2$. This coefficient is equal to the squared coefficient of correlation between the dependent variable (Y) and the predicted dependent variable ($\hat{Y}$).

An alternative way of computing the multiple coefficient of correlation is to divide the regression sum of squares by the total sum of squares. This

shows that $R_{Y.1, \dots, J}^2$ can also be interpreted as the proportion of variance of the dependent variable explained by the independent variables. With this interpretation, the multiple coefficient of correlation is computed as

$$R_{Y.1, \dots, J}^2 = \frac{SS_{\text{regression}}}{SS_{\text{regression}} + SS_{\text{error}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}}. \quad (7)$$

Significance Test

In order to assess the significance of a given $R_{Y.1, \dots, J}^2$, we can compute an F ratio as

$$F = \frac{R_{Y.1, \dots, J}^2}{1 - R_{Y.1, \dots, J}^2} \times \frac{N - J - 1}{J}. \quad (8)$$

Under the usual assumptions of normality of the error and of independence of the error and the scores, this F ratio is distributed under the null hypothesis as a Fisher distribution with $v_1 = J$ and $v_2 = N - J - 1$ degrees of freedom.

Estimating the Population Correlation: Shrunk and Adjusted R

Just like its bivariate counterpart r , the multiple coefficient of correlation is a *descriptive* statistic that always overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. In order to obtain a better estimate of the population, the value $R_{Y.1, \dots, J}^2$ needs to be corrected. The corrected value of $R_{Y.1, \dots, J}^2$ goes by different names: *corrected R* , *shrunk R* , or *adjusted R* (there are some subtle differences between these different appellations, but we will ignore them here), and we denote it by $\tilde{R}_{Y.1, \dots, J}^2$. There are several correction formulas available; the one most often used estimates the value of the population correlation as

$$\tilde{R}_{Y.1, \dots, J}^2 = 1 - \left[(1 - R_{Y.1, \dots, J}^2) \left(\frac{N - 1}{N - J - 1} \right) \right]. \quad (9)$$

Example 1: Multiple Correlation Coefficient With Orthogonal Predictors

When the independent variables are pairwise orthogonal, the importance of each of them in the regression is assessed by computing the squared coefficient of correlation between each of the independent variables and the dependent variable. The sum of these squared coefficients of correlation is equal to the multiple coefficient of correlation. We illustrate this case with the data from Table 1. In this example, the dependent variable (Y) is the number of sentences recalled by participants who learned a list of unrelated sentences. The first independent variable or first predictor, X_1 , is the number of trials used to learn the list. It takes the values 2, 4, and 8. It is expected that recall will increase as a function of the number of trials. The second independent variable, X_2 , is the number of additional interpolated lists that the participants are asked to learn. It takes the values 2, 4, and 8. As a consequence of retroactive inhibition, it is expected that recall will decrease as a function of the number of interpolated lists learned.

Using Equation 3, we found that $\hat{Y}$ can be obtained from X_1 and X_2 as

$$\hat{Y} = 30 + 6 \times X_1 - 4 \times X_2. \quad (10)$$

Table 1 A Set of Data

Number of Learning Trials (X)	Number of Interpolated Lists (T)		
	2	4	8
2	35	21	6
	39	31	8
4	40	34	18
	52	42	26
8	61	58	46
	73	66	52

Source: Data from Abdi et al. (2002).

Notes: The dependent variable Y is to be predicted from two orthogonal predictors X_1 and X_2 . These data are the results of a hypothetical experiment on retroactive interference and learning. Y is the number of sentences remembered from a set of sentences learned, X_1 is the number of learning trials, and X_2 is the number of interpolated lists learned.

Using these data and Equations 4 and 5, we find that

$$SS_{\text{regression}} = 5824, SS_{\text{total}} = 6214, \text{ and } SS_{\text{error}} = 390. \quad (11)$$

This gives the following value for the multiple coefficient of correlation:

$$R_{Y.1, \dots, J}^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{5824}{6214} = .9372. \quad (12)$$

In order to decide if this value of $R_{Y.1, \dots, J}^2$ is large enough to be considered significant, we compute an F ratio equal to

$$\begin{aligned} F &= \frac{R_{Y.1, \dots, J}^2}{1 - R_{Y.1, \dots, J}^2} \times \frac{N - J - 1}{J} \\ &= \frac{.9372}{1 - .9372} \times \frac{15}{2} = 111.93. \end{aligned} \quad (13)$$

Such a value of F is significant at all the usual alpha levels, and therefore we can reject the null hypothesis.

Because X_1 and X_2 are orthogonal to each other (i.e., their correlation is equal to 0), the multiple coefficient of correlation is equal to the sum of the squared coefficients of correlation between the independent variables and the dependent variable:

$$R_{Y.1, \dots, J}^2 = .9372 = r_{Y.1}^2 + r_{Y.2}^2 = .6488 + .2884. \quad (14)$$

A better estimate of the population value of the multiple coefficient of correlation can be obtained as

$$\begin{aligned} \tilde{R}_{Y.1, \dots, J}^2 &= 1 - \left[(1 - R_{Y.1, \dots, J}^2) \left(\frac{N - 1}{N - J - 1} \right) \right] \\ &= 1 - (1 - .9372) \frac{17}{15} = .9289. \end{aligned} \quad (15)$$

Example 2: Multiple Correlation Coefficient With Nonorthogonal Predictors

When the independent variables are correlated, the multiple coefficient of correlation is not equal to the

sum of the squared correlation coefficients between the dependent variable and the independent variables. In fact, such a strategy would *overestimate* the contribution of each variable because the variance that they share would be counted several times.

For example, consider the data given in Table 2, where the dependent variable is to be predicted from the independent variables X_1 and X_2 . The prediction of the dependent variable (using Equation 3) is found to be equal to

$$\hat{Y} = 1.67 + X_1 + 9.50X_2, \quad (16)$$

which gives a multiple coefficient of correlation of $R_{Y.1, \dots, J}^2 = .9866$. The coefficient of correlation between X_1 and X_2 is equal to $r_{X_1, X_2} = .7500$, between X_1 and Y is equal to $r_{Y.1} = .8028$, and between X_2 and Y is equal to $r_{Y.2} = .9890$. It can easily be checked that the multiple coefficient of correlation is not equal to the sum of the squared coefficients of correlation between the independent variables and the dependent variables:

$$\begin{aligned} R_{Y.1, \dots, J}^2 &= .9866 \\ &\neq r_{Y.1}^2 + r_{Y.2}^2 \\ &= .665 + .9780 = 1.6225. \end{aligned} \quad (17)$$

Using the data from Table 2 along with Equations 4 and 5, we find that

$$\begin{aligned} SS_{\text{regression}} &= 1822.00, \\ SS_{\text{total}} &= 1846.83, \text{ and} \\ SS_{\text{error}} &= 24.83. \end{aligned} \quad (18)$$

Table 2 A Set of Data

Y (Memory span)	14	23	30	50	39	67
X_1 (Age)	4	4	7	7	10	10
X_2 (Speech rate)	1	2	2	4	3	6

Source: Data from Abdi et al. (2002).

Notes: The dependent variable Y is to be predicted from two correlated (i.e., nonorthogonal) predictors: X_1 and X_2 . Y is the number of digits a child can remember for a short time (the “memory span”), X_1 is the age of the child, and X_2 is the speech rate of the child (how many words the child can pronounce in a given time). Six children were tested.

This gives the following value for the multiple coefficient of correlation:

$$R_{Y.1, \dots, J}^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{1822.00}{1846.83} = .9866. \quad (19)$$

In order to decide if this value of $R_{Y.1, \dots, J}^2$ is large enough to be considered significant, we compute an F ratio equal to

$$\begin{aligned} F &= \frac{R_{Y.1, \dots, J}^2}{1 - R_{Y.1, \dots, J}^2} \times \frac{N - J - 1}{J} \\ &= \frac{.9866}{1 - .9866} \times \frac{3}{2} = 110.50. \end{aligned} \quad (20)$$

Such a value of F is significant at all the usual alpha levels, and therefore we can reject the null hypothesis.

A better estimate of the population value of the multiple coefficient of correlation can be obtained as

$$\begin{aligned} \tilde{R}_{Y.1, \dots, J}^2 &= 1 - \left[(1 - R_{Y.1, \dots, J}^2) \left(\frac{N - 1}{N - J - 1} \right) \right] \\ &= 1 - (1 - .9866) \frac{5}{2} = .9776. \end{aligned} \quad (21)$$

—Hervé Abdi

See also Coefficients of Correlation, Alienation, and Determination; Correlation Coefficient

Further Reading

- Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

MULTIPLE CORRESPONDENCE ANALYSIS

Multiple correspondence analysis (MCA) is an extension of correspondence analysis (CA), which allows one to analyze the pattern of relationships of several categorical dependent variables. As such, it can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative. Because MCA has been (re)discovered many times, equivalent methods are known under several different names, such as optimal scaling, optimal or appropriate scoring, dual scaling, homogeneity analysis, scalogram analysis, and quantification method.

Technically, MCA is obtained by using a standard correspondence analysis on an indicator matrix (i.e., a matrix whose entries are 0 or 1). The percentages of explained variance need to be corrected, and the correspondence analysis interpretation of inter-point distances needs to be adapted.

When to Use It

MCA is used to analyze a set of observations described by a set of nominal variables. Each nominal variable is composed of several levels, and each of these levels is coded as a binary variable. For example, gender (F vs. M) is one nominal variable with two levels. The pattern for a male respondent will be 01, and 10 for a female. The complete data table is composed of binary columns with one and only one column taking the value “1” per nominal variable.

MCA can also accommodate quantitative variables by recoding them as “bins.” For example, a score with a range of -5 to $+5$ could be recoded as a nominal variable with three levels: less than 0, equal to 0, or more than 0. With this schema, a value of 3 will be expressed by the pattern 001. The coding schema of MCA implies that each row has the same total, which for CA implies that each row has the same *mass*.

An Example

We illustrate the method with an example from wine testing. Suppose that we want to evaluate the effect of the oak species on barrel aged red Burgundy wines. First, we aged wine coming from the same harvest of Pinot Noir in six different barrels made with two types of oak. Wines 1, 5, and 6 were aged with the first type of oak, whereas wines 2, 3, and 4 were aged with the second. Next, we asked each of three wine experts to choose from two to five variables to describe the wines. For each wine and for each variable, the expert was asked to rate the intensity. The answer given by the expert was coded either as a binary answer (i.e., fruity vs. non fruity) or as a ternary answer (i.e., no vanilla, a bit of vanilla, clear smell of vanilla). Each binary answer is represented by 2 binary columns (e.g., the answer “fruity” is represented by the pattern 10 and “nonfruity” is 01). A ternary answer is represented by 3 binary columns (i.e., the answer “some vanilla” is represented by the pattern 010). The results are presented in Table 1. The goal of the analysis is twofold. First, we want to obtain a typology of the wines, and second, we want to know if there is an agreement between the scales used by the experts. We will use the type of oak as a supplementary (or illustrative) variable to be projected on the analysis after the fact. Also, after the testing of the six wines was performed, an unknown bottle of Pinot Noir was found and tested by the wine testers. This wine will be used as a supplementary observation. For this wine,

when an expert was not sure of how to use a descriptor, a pattern of response such as .5.5 was used to represent the answer.

Notations

There are K nominal variables, each nominal variable has J_k levels, and the sum of the J_k levels is equal to J . There are I observations. The $I \times J$ indicator matrix is denoted $\mathbf{X}$. Performing CA on the indicator matrix will provide two sets of factor scores: one for the rows and one for the columns. These factor scores are, in general, scaled such that their variance is equal to their corresponding eigenvalue (some versions of CA compute row factor scores normalized to unity).

The grand total of the table is noted N , and the first step of the analysis is to compute the probability matrix $\mathbf{Z} = N^{-1}\mathbf{X}$. We denote $\mathbf{r}$ the vector of the row totals of $\mathbf{Z}$ (i.e., $\mathbf{r} = \mathbf{Z}\mathbf{1}$, with $\mathbf{1}$ being a conformable vector of 1s), $\mathbf{c}$ the vector of the column totals, and $\mathbf{D}_c = \text{diag}\{\mathbf{c}\}$, $\mathbf{D}_r = \text{diag}\{\mathbf{r}\}$. The factor scores are obtained from the following singular value decomposition:

$$\mathbf{D}_r^{-\frac{1}{2}}(\mathbf{Z} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{P}\mathbf{A}\mathbf{Q}^T. \quad (1)$$

Δ is the diagonal matrix of the *singular* values, and $\mathbf{\Lambda} = \Delta^2$ is the matrix of the eigenvalues. The row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{P}\mathbf{\Lambda} \quad \text{and} \quad \mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{Q}\mathbf{\Lambda}. \quad (2)$$

Table 1 Data for the Barrel-Aged Red Burgundy Wines Example

Wine	Oak Type	Expert 1						Expert 2						Expert 3		
		Fruity		Woody		Coffee		Red Fruit		Roasted		Vanillin		Woody		
W1	1	1	0	0	0	1	0	1	0	0	1	0	0	1	0	1
W2	2	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0
W3	2	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0
W4	2	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0
W5	1	1	0	0	0	1	0	1	1	0	0	0	1	0	1	0
W6	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0	1
W?	?	0	1	0	1	0	.5	.5	1	0	1	0	0	1	.5	.5

Notes: “Oak Type” is an illustrative (supplementary) variable, The wine W? is an unknown wine treated as a supplementary observation.

The squared (χ^2) distances from the rows and columns to their respective barycenters are obtained as

$$\mathbf{d}_r = \text{diag}\{\mathbf{F}\mathbf{F}^T\} \quad \text{and} \quad \mathbf{d}_c = \text{diag}\{\mathbf{G}\mathbf{G}^T\}. \quad (3)$$

The squared *cosines* between row i and factor ℓ and column j and factor ℓ are obtained respectively as

$$o_{i,\ell} = \frac{f_{i,\ell}^2}{d_{r,i}^2} \quad \text{and} \quad o_{j,\ell} = \frac{g_{j,\ell}^2}{d_{c,j}^2} \quad (4)$$

(with $d_{r,i}^2$ and $d_{c,j}^2$ being, respectively, the i th element of $\mathbf{d}_r$ and the j th element of $\mathbf{d}_c$). Squared cosines help locate the factors important for a given observation or variable. The *contribution* of row i to factor ℓ and of column j to factor ℓ are obtained respectively as

$$t_{i,\ell} = \frac{f_{i,\ell}^2}{\lambda_\ell} \quad \text{and} \quad t_{j,\ell} = \frac{g_{j,\ell}^2}{\lambda_\ell}. \quad (5)$$

Contributions help locate the observations or variables important for a given factor.

Supplementary or illustrative elements can be projected onto the factors using the so called *transition* formula. Specifically, let $\mathbf{i}_{\text{sup}}^T$ be an illustrative row and $\mathbf{j}_{\text{sup}}$ be an illustrative column to be projected. Their coordinates $\mathbf{f}_{\text{sup}}$ and $\mathbf{g}_{\text{sup}}$ are obtained as

$$\mathbf{f}_{\text{sup}} = (\mathbf{i}_{\text{sup}}^T \mathbf{1}) \mathbf{i}_{\text{sup}}^T \mathbf{G} \mathbf{\Delta}^{-1} \quad \text{and} \quad \mathbf{g}_{\text{sup}} = (\mathbf{j}_{\text{sup}}^T \mathbf{1}) \mathbf{j}_{\text{sup}}^T \mathbf{F} \mathbf{\Delta}^{-1}. \quad (6)$$

Performing CA on the indicator matrix will provide factor scores for the rows and the columns. However, the factor scores given by a CA program will need to be rescaled for MCA, as explained in the next section.

The $J \times J$ table obtained as $\mathbf{B} = \mathbf{X}^T \mathbf{X}$ is called the *Burt matrix* associated to $\mathbf{X}$. This table is important in MCA because using CA on the Burt matrix gives the same factors as the analysis of $\mathbf{X}$ but is often computationally easier. But the Burt matrix also plays an important theoretical role because the eigenvalues obtained from its analysis give a better approximation of the inertia explained by the factors than do the eigenvalues of $\mathbf{X}$.

Eigenvalue Correction for Multiple Correspondence Analysis

MCA codes data by creating several binary columns for each variable with the constraint that one and only one of the columns gets the value 1. This coding schema creates artificial additional dimensions because *one* categorical variable is coded with *several* columns. As a consequence, the inertia (i.e., variance) of the solution space is artificially inflated, and therefore, the percentage of inertia explained by the first dimension is severely *underestimated*. In fact, it can be shown that all the factors with an eigenvalue less or equal to $\frac{1}{K}$ simply code these additional dimensions ($K = 10$ in our example).

Two correction formulas are often used; the first one is due to Benzécri, and the second one to Greenacre. These formulas take into account that the eigenvalues smaller than $\frac{1}{K}$ are coding for the extra dimensions and that MCA is equivalent to the analysis of the Burt matrix, whose eigenvalues are equal to the squared eigenvalues of the analysis of $\mathbf{X}$. Specifically, if we denote by λ_ℓ the eigenvalues obtained from the analysis of the indicator matrix, then the corrected eigenvalues, denoted $c\lambda_\ell$, are obtained as

$$c\lambda_\ell = \begin{cases} \left[\left(\frac{K}{K-1} \right) \left(\lambda_\ell - \frac{1}{K} \right) \right]^2 & \text{if } \lambda_\ell > \frac{1}{K} \\ 0 & \text{if } \lambda_\ell \leq \frac{1}{K} \end{cases} \quad (7)$$

Using this formula gives a better estimate of the inertia extracted by each eigenvalue.

Traditionally, the percentages of inertia are computed by dividing each eigenvalue by the sum of the eigenvalues, and this approach could be used here also. However, it will give an *optimistic* estimation of the percentage of inertia. A better estimation of the inertia has been proposed by Greenacre, who suggested instead to evaluate the percentage of inertia relative to the average inertia of the off diagonal blocks of the Burt matrix. This average inertia, denoted $\bar{\mathcal{I}}$, can be computed as

$$\bar{\mathcal{I}} = \frac{K}{K-1} \times \left(\sum_{\ell} \lambda_\ell^2 - \frac{J-K}{K} \right)^2. \quad (8)$$

According to this approach, the percentage of inertia would be obtained by the ratio

$$\tau_c = \frac{c\lambda}{\bar{I}} \text{ instead of } \frac{c\lambda}{\sum c\lambda_\ell}. \quad (9)$$

Interpreting MCA

As with CA, the interpretation in MCA is often based upon proximities between points in a low-dimensional map (i.e., two or three dimensions). As well as for CA, proximities are meaningful only between points from the same set (i.e., rows with rows, columns with columns). Specifically, when two row points are close to each other, they tend to select the same levels of the

nominal variables. For the proximity between variables, we need to distinguish two cases. First, the proximity between levels of *different* nominal variables means that these levels tend to appear together in the observations. Second, because the levels of the *same* nominal variable cannot occur together, we need a different type of interpretation for this case. Here, the proximity between levels means that the groups of observations associated with these two levels are themselves similar.

The Example

Table 2 lists the corrected eigenvalues and proportion of explained inertia obtained with the Benzécri/Greenacre correction formula. Tables 3 and 4

Table 2 Eigenvalues, Corrected Eigenvalues, Proportion of Explained Inertia, and Corrected Proportion of Explained Inertia

Factor	Indicator Matrix		Burt Matrix		Benzécri Correction		Greenacre Correction	
	λ	τ_i	${}_B\lambda$	τ_B	${}_z\lambda$	τ_z	${}_c\lambda$	τ_c
1	.8532	.7110	.7280	.9306	.7004	.9823	.7004	.5837
2	.2000	.1667	.0400	.0511	.0123	.0173	.0123	.0103
3	.1151	.0959	.0133	.0169	.0003	.0004	.0003	.0002
4	.0317	.0264	.0010	.0013	0	0	0	0
Σ	1.2000	1	.7822	1	.7130	1	.7130	.5942

Notes: The eigenvalues of the Burt matrix are equal to the squared eigenvalues of the indicator matrix. The corrected eigenvalues for Benzécri and Greenacre are the same, but the proportion of explained variance differs. Eigenvalues are denoted by λ , proportions of explained inertia by τ .

Table 3 Factor Scores, Squared Cosines, and Contributions for the Observations (*I* set)

<i>F</i>	${}_c\lambda$	%c	Wine 1	Wine 2	Wine 3	Wine 4	Wine 5	Wine 6	Wine ?
			Factor Scores						
1	.7004	58	0.86	−0.71	−0.92	−0.86	0.92	0.71	0.03
2	.0103	1	0.08	−0.16	0.08	0.08	0.08	−0.16	−0.16
<i>F</i>	Squared Cosines								
1			.62	.42	.71	.62	.71	.42	.04
2			.01	.02	.01	.01	.01	.02	.96
<i>F</i>	Contributions × 1000								
1			177	121	202	177	202	121	—
2			83	333	83	83	83	333	—

Notes: The eigenvalues and proportions of explained inertia are corrected using the Benzécri/Greenacre formula. Contributions corresponding to negative scores are in *italic*. The mystery wine (Wine ?) is a supplementary observation. Only the first two factors are reported.

Table 4 Factor Scores, Squared Cosines, and Contributions for the Variables (*I* set)

Expert 1										Expert 2						Expert 3																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
Fruity					Woody					Red					Fruit			Roasted			Vanillin			Woody			Fruity			Butter			Woody			Oak																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
y					n					y					n			y			n			y			n			y			n			y		n																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											
F					λ					$\%_c$					Factor Scores																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																		
1					.7004					58					.90					-.90			.90			-.97			.00			.97			-.90			.90			.28		-.28		-.90			.90			.90		-.90																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
2					.0103					1					.00					.00			.00			.18			-.35			.18			.00			.00			.00			.00			.00		.00		.00		.00																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
F															Squared					Cosines					Contributions																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																								
																														$\times 1000$																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
1										58					58			58			58			44			0			44			58			58			58			58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58		58	

Notes: The eigenvalues and percentages of inertia have been corrected using the Benzécri/Greenacre formula. Contributions corresponding to negative scores are in italic. Oak 1 and 2 are supplementary variables.

give the corrected factor scores, cosines, and contributions for the rows and columns of Table 1. Figure 1 displays the projections of the rows and the columns. We have separated these two sets, but because the projections have the same variance, these two graphs could be displayed together (as long as one keeps in mind that distances between points are meaningful only within the same set). The analysis is essentially unidimensional, with wines 2, 3, and 4 clustered on the negative side of the factors and wines 1, 5, and 6 on the positive side. The supplementary wine does not seem to belong to either cluster. The analysis of the columns shows that the negative side of the factor is characterized as being nonfruity, nonwoody, and coffee by Expert 1; roasted, nonfruity, low in vanilla, and woody by Expert 2; and buttery and woody by Expert 3. The positive side gives the reverse pattern. The supplementary elements indicate that the negative side is correlated with the second type of oak, whereas the positive side is correlated with the first type of oak.

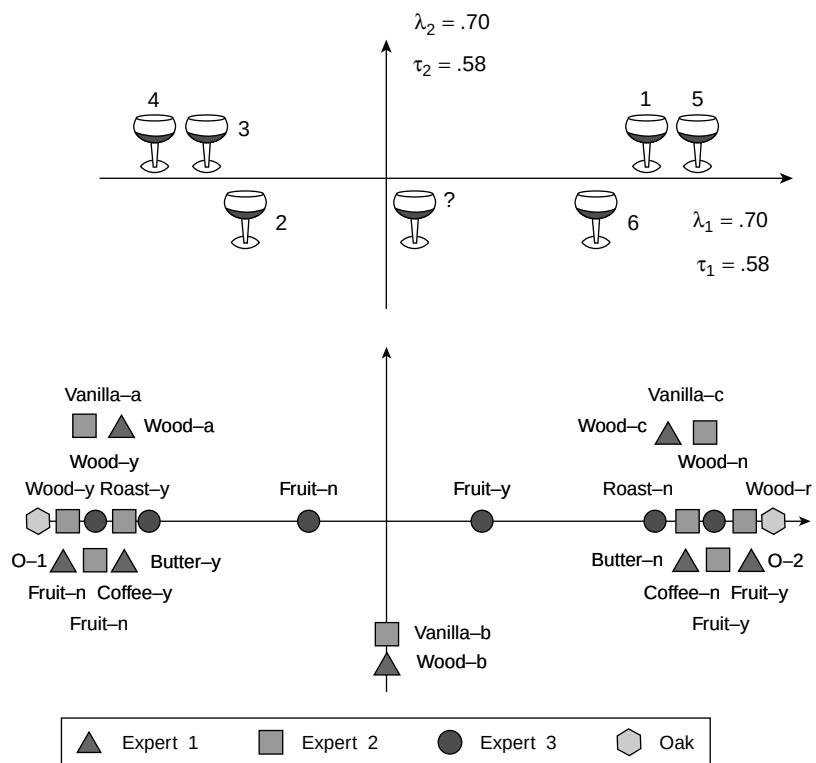


Figure 1 Multiple Correspondence Analysis—Projections on the First Two Dimensions

Notes: The eigenvalues (λ) and proportion of explained inertia (τ) have been corrected with the Benzécri/Greenacre formula. (a) The I set: rows (i.e., wines). Wine ? is a supplementary element. (b) The J set: columns (i.e., adjectives). Oak 1 and Oak 2 are supplementary elements (the projection points have been slightly moved to increase readability). (Projections from Tables 3 and 4.)

Alternatives to MCA

Because the interpretation of MCA is more delicate than simple CA, several approaches have been suggested to offer the simplicity of interpretation of CA for indicator matrices. One approach is to use a different metric from χ^2 , the most attractive alternative being the Hellinger distance. Another approach, called *joint correspondence analysis*, fits only the off-diagonal tables of the Burt matrix and can be interpreted as a factor analytic model.

—Hervé Abdi and Dominique Valentin

See also Centroid; Correspondence Analysis; Distance; DISTATIS; Eigendecomposition; Metric Multidimensional Scaling; Multiple Factor Analysis; Singular and Generalized Singular Value Decomposition; STATIS

Further Reading

- Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 4, 377–378.
- Clausen, S. E. (1998). *Applied correspondence analysis*. Thousand Oaks, CA: Sage.
- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistiques Appliquées*, 26, 29–37.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.
- Rao, C. (1995). The use of Hellinger distance in graphical displays of contingency table data. In E. M. Tiit, T. Kollo, &

H. Niemi (Eds.), *Multivariate statistics and matrices in statistics* (Vol. 3, pp. 143–161). Leiden, Netherlands: Brill.
 Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Newbury Park, CA: Sage.

MULTIPLE FACTOR ANALYSIS

Multiple factor analysis (MFA) analyzes observations described by several “blocks” or sets of variables. MFA seeks the common structures present in all or some of these sets. MFA is performed in two steps. First, a principal component analysis (PCA) is performed on each data set, which is then “normalized” by dividing all its elements by the square root of the first eigenvalue obtained from its PCA. Second, the normalized data sets are merged to form a unique matrix, and a global PCA is performed on this matrix. The individual data sets are then projected onto the global analysis to analyze communalities and discrepancies. MFA is used in very different domains such as sensory evaluation, economy, ecology, and chemistry.

MFA is used to analyze a set of observations described by several groups of variables. The number of variables in each group may differ, and the nature of the variables (nominal or quantitative) can vary from one group to the other, but the variables should be of the same nature in a given group. The analysis derives an integrated picture of the observations and of the relationships between the groups of variables.

The goal of MFA is to integrate different groups of variables describing the same observations. In order to do so, the first step is to make these groups of variables

comparable. Such a step is needed because the straightforward analysis obtained by concatenating all variables would be dominated by the group with the strongest structure. A similar problem can occur in a non-normalized PCA: Without normalization, the structure is dominated by the variables with the largest variance. For PCA, the solution is to normalize (i.e., to use *z* scores) each variable by dividing it by its standard deviation. The solution proposed by MFA is similar: To compare groups of variables, each group is normalized by dividing all its elements by a quantity called its *first singular value*, which is the matrix equivalent of the standard deviation. Practically, this step is implemented by performing a PCA on each group of variables. The first singular value is the square root of the first eigenvalue of the PCA. After normalization, the data tables are concatenated into a data table that is submitted to PCA.

An Example

To illustrate MFA, we selected six wines, coming from the same harvest of Pinot Noir, aged in six different barrels made with one of two different types of oak. Wines 1, 5, and 6 were aged with the first type of oak, and wines 2, 3, and 4 with the second. Next, we asked each of three wine experts to choose from two to five variables to describe the six wines. For each wine, the expert rated the intensity of the variables on a 9 point scale. The results are presented in Table 1 (the same example is used in the entry for STATIS). The goal of the analysis is twofold. First, we want to obtain a typology of the wines, and second, we want to know if there is an agreement between the experts.

Table 1 Raw Data for the Wine Example

Wines	Oak Type	Expert 1			Expert 2				Expert 3		
		Fruity	Woody	Coffee	Red Fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
Wine 1	1	1	6	7	2	5	7	6	3	6	7
Wine 2	2	5	3	2	4	4	4	2	4	4	3
Wine 3	2	6	1	1	5	2	1	1	7	1	1
Wine 4	2	7	1	2	7	2	1	2	2	2	2
Wine 5	1	2	5	4	3	5	6	5	2	6	6
Wine 6	1	3	4	4	3	5	4	5	1	7	5

Notations

The raw data consist in T data sets. Each data set is called a *study*. Each study is an $I \times J_{[t]}$ rectangular data matrix denoted $\mathbf{Y}_{[t]}$, where I is the number of observations and $J_{[t]}$ the number of variables of the t th study. Each data matrix is, in general, preprocessed (e.g., centered, normalized), and the preprocessed data matrices actually used in the analysis are denoted $\mathbf{X}_{[t]}$.

For our example, the data consist in $T = 3$ studies. The data (from Table 1) were centered by column (i.e., the mean of each column is zero) and normalized (i.e., for each column, the sum of the squared elements is equal to 1). So the starting point of the analysis consists in three matrices:

$$\begin{aligned} \mathbf{X}_{[1]} &= \begin{bmatrix} -0.57 & 0.58 & 0.76 \\ 0.19 & -0.07 & -0.28 \\ 0.38 & -0.50 & -0.48 \\ 0.57 & -0.50 & -0.28 \\ -0.38 & 0.36 & 0.14 \\ -0.19 & 0.14 & 0.14 \end{bmatrix}, \\ \mathbf{X}_{[2]} &= \begin{bmatrix} -0.50 & 0.35 & 0.57 & 0.54 \\ 0.00 & 0.05 & 0.03 & -0.32 \\ 0.25 & -0.56 & -0.51 & -0.54 \\ 0.75 & -0.56 & -0.51 & -0.32 \\ -0.25 & 0.35 & 0.39 & 0.32 \\ -0.25 & 0.35 & 0.03 & 0.32 \end{bmatrix} \\ \text{and } \mathbf{X}_{[3]} &= \begin{bmatrix} -0.03 & 0.31 & 0.57 \\ 0.17 & -0.06 & -0.19 \\ 0.80 & -0.61 & -0.57 \\ -0.24 & -0.43 & -0.38 \\ -0.24 & 0.31 & 0.38 \\ -0.45 & 0.49 & 0.19 \end{bmatrix}. \end{aligned} \quad (1)$$

Each observation is assigned a *mass* that reflects its importance. When all observations have the same importance, their masses are all equal to $m_i = \frac{1}{I}$. The set of the masses is stored in an $I \times I$ diagonal matrix denoted $\mathbf{M}$.

Finding the Global Space

Computing the Separate PCAs

To normalize the studies, we first compute a PCA for each study. The first singular value (i.e., the square root of the first eigenvalue) is the normalizing factor used to divide the elements of the data table. For example, the PCA of the first group gives a first eigenvalue ${}_1\varphi_1 = 2.86$ and a first singular value of ${}_1\varphi_1 = \sqrt{{}_1\varphi_1} = 1.69$. This gives the first normalized data matrix, denoted $\mathbf{Z}_{[1]}$:

$$\mathbf{Z}_{[1]} = {}_1\varphi_1^{-1} \times \mathbf{X}_{[1]} = \begin{bmatrix} -0.33 & 0.34 & 0.45 \\ 0.11 & -0.04 & -0.16 \\ 0.22 & -0.30 & -0.28 \\ 0.33 & -0.30 & -0.16 \\ -0.22 & 0.21 & 0.08 \\ -0.11 & 0.08 & 0.08 \end{bmatrix}. \quad (2)$$

Matrices $\mathbf{Z}_{[2]}$ and $\mathbf{Z}_{[3]}$ are normalized with their first respective singular values of ${}_2\varphi_1 = 1.91$ and ${}_3\varphi_1 = 1.58$. Normalized matrices have a first singular value equal to 1.

Building the Global Matrix

The normalized studies are concatenated into an $I \times T$ matrix called the *global data matrix*, denoted $\mathbf{Z}$. Here we obtain

$$\mathbf{Z} = [\mathbf{Z}_{[1]} \mathbf{Z}_{[2]} \mathbf{Z}_{[3]}] = \begin{bmatrix} -0.33 & 0.34 & 0.45 & -0.26 & 0.18 & 0.30 & 0.28 & -0.02 & 0.19 & 0.36 \\ 0.11 & -0.04 & -0.16 & 0.00 & 0.03 & 0.02 & -0.17 & 0.11 & -0.04 & -0.12 \\ 0.22 & -0.30 & -0.28 & 0.13 & -0.29 & -0.27 & -0.28 & 0.51 & -0.39 & -0.36 \\ 0.33 & -0.30 & -0.16 & 0.39 & -0.29 & -0.27 & -0.17 & -0.15 & -0.27 & -0.24 \\ -0.22 & 0.21 & 0.08 & -0.13 & 0.18 & 0.20 & 0.17 & -0.15 & 0.19 & 0.24 \\ -0.11 & 0.08 & 0.08 & -0.13 & 0.18 & 0.02 & 0.17 & -0.29 & 0.31 & 0.12 \end{bmatrix}. \quad (3)$$

Computing the Global PCA

To analyze the global matrix, we use standard PCA. This amounts to computing the singular value decomposition of the global data matrix

$$\mathbf{Z} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \quad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}, \quad (4)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{Z}$, and $\mathbf{\Delta}$ is the diagonal matrix of the singular values.

For our example, we obtain

$$\mathbf{U} = \begin{bmatrix} 0.53 & -0.35 & -0.58 & -0.04 & 0.31 \\ -0.13 & -0.13 & 0.49 & 0.51 & 0.54 \\ -0.56 & -0.57 & 0.01 & -0.36 & -0.25 \\ -0.44 & 0.62 & -0.48 & 0.15 & 0.03 \\ 0.34 & 0.04 & 0.16 & 0.39 & -0.73 \\ 0.27 & 0.40 & 0.40 & -0.65 & 0.11 \end{bmatrix} \quad (5)$$

and

$$\text{diag}\{\mathbf{\Delta}\} = [1.68 \quad 0.60 \quad 0.34 \quad 0.18 \quad 0.11]$$

and

$$\begin{aligned} \text{diag}\{\mathbf{\Lambda}\} &= \text{diag}\{\mathbf{\Delta}^2\} \\ &= [2.83 \quad 0.36 \quad 0.11 \quad 0.03 \quad 0.01] \end{aligned} \quad (6)$$

($\mathbf{\Lambda}$ gives the eigenvalues of the PCA) and

$$\mathbf{V} = \begin{bmatrix} -0.34 & 0.22 & 0.03 & 0.14 & 0.55 \\ 0.35 & -0.14 & -0.03 & 0.30 & 0.02 \\ 0.32 & -0.06 & -0.65 & -0.24 & 0.60 \\ -0.28 & 0.34 & -0.32 & 0.31 & -0.18 \\ 0.30 & -0.00 & 0.43 & 0.11 & 0.19 \\ 0.30 & -0.18 & -0.00 & 0.67 & 0.11 \\ 0.30 & 0.09 & -0.22 & -0.36 & -0.38 \\ -0.22 & -0.86 & 0.01 & -0.12 & -0.00 \\ 0.36 & 0.20 & 0.45 & -0.30 & 0.19 \\ 0.37 & 0.01 & -0.21 & 0.18 & -0.28 \end{bmatrix}. \quad (7)$$

The global factor scores for the wines are obtained as

$$\mathbf{F} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta} \quad (8)$$

$$= \begin{bmatrix} 2.18 & -0.51 & -0.48 & -0.02 & 0.08 \\ -0.56 & -0.20 & 0.41 & 0.23 & 0.15 \\ -2.32 & -0.83 & 0.01 & -0.16 & -0.07 \\ -1.83 & 0.90 & -0.40 & 0.07 & 0.01 \\ 1.40 & 0.05 & 0.13 & 0.17 & -0.20 \\ 1.13 & 0.58 & 0.34 & -0.29 & 0.03 \end{bmatrix}. \quad (9)$$

In $\mathbf{F}$, each row represents an observation (i.e., a wine) and each column a component. Figure 1 displays the wines in the space of the first two principal components. The first component has an eigenvalue equal to $\lambda_1 = 2.83$, which corresponds to 84% of the inertia ($\frac{2.83}{2.83+0.36+0.11+0.03+0.01} = \frac{2.83}{3.35} \approx .84$). The second component, with an eigenvalue of .36, explains 11% of the inertia. The first component is interpreted as the opposition between the first (wines 1, 5, and 6) and the second oak type (wines 2, 3, and 4).

Partial Analyses

The global analysis reveals the common structure of the wine space. In addition, we want to see how each expert “interprets” this space. This is achieved by projecting the data set of each expert onto the global analysis. This is implemented by multiplication of a cross product matrix by a projection matrix. The projection matrix is obtained by

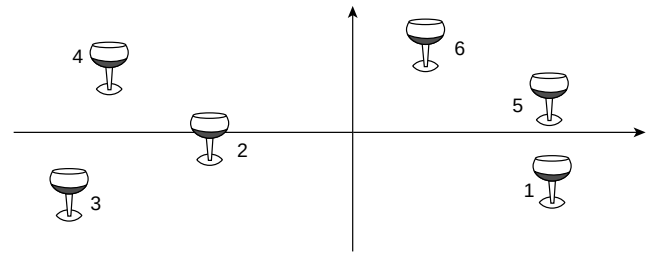


Figure 1 Global Analysis: Plot of the Wines on the First Two Principal Components

Notes: First component: $\lambda_1 = 2.83$, explains 84% of the inertia; second component: $\lambda_2 = 0.36$, explains 11% of the inertia.

rewriting Equation 8 to show that the global factor scores could be computed as

$$\mathbf{F} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta} = (\mathbf{Z}\mathbf{Z}^T) \times (\mathbf{M}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta}^{-1}). \quad (10)$$

This shows that $\mathbf{P} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta}^{-1}$ is a projection matrix that transforms the matrix $\mathbf{Z}\mathbf{Z}^T$ into factor scores. Here, we obtain

$$\mathbf{P} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta}^{-1} = \begin{bmatrix} 0.77 & -1.43 & -4.20 & -0.55 & 6.68 \\ -0.20 & -0.55 & 3.56 & 6.90 & 11.71 \\ -0.82 & -2.33 & 0.05 & -4.85 & -5.53 \\ -0.65 & 2.54 & -3.46 & 2.01 & 0.66 \\ 0.50 & 0.15 & 1.13 & 5.27 & -15.93 \\ 0.40 & 1.61 & 2.91 & -8.78 & 2.43 \end{bmatrix}, \quad (11)$$

which is then used to project the studies onto the global space. For example, for the first expert we obtain

$$\mathbf{F}_{[1]} = T \times (\mathbf{Z}_{[1]}\mathbf{Z}_{[1]}^T)\mathbf{P} = \begin{bmatrix} 2.76 & -1.10 & -2.29 & -0.39 & 0.67 \\ -0.77 & 0.30 & 0.81 & 0.31 & -0.27 \\ -1.99 & 0.81 & 1.48 & 0.08 & -0.39 \\ -1.98 & 0.93 & 0.92 & -0.02 & 0.59 \\ 1.29 & -0.62 & -0.49 & 0.10 & -0.51 \\ 0.69 & -0.30 & -0.43 & -0.07 & -0.08 \end{bmatrix}. \quad (12)$$

(Multiplying by T is needed in order to scale one expert with all $T = 3$ experts of the global solution.) The same procedure is used for Experts 2 and 3:

$$\mathbf{F}_{[2]} = \begin{bmatrix} 2.21 & -0.86 & 0.74 & 0.27 & 0.06 \\ -0.28 & -0.13 & 0.35 & 0.55 & 0.52 \\ -2.11 & 0.50 & -0.77 & -0.49 & -0.01 \\ -2.39 & 1.23 & -1.57 & -0.20 & -0.68 \\ 1.49 & -0.49 & 0.62 & 0.40 & 0.13 \\ 1.08 & -0.24 & 0.63 & -0.53 & -0.03 \end{bmatrix}, \quad (14)$$

and

$$\mathbf{F}_{[3]} = \begin{bmatrix} 1.54 & 0.44 & 0.09 & 0.07 & -0.47 \\ -0.61 & -0.76 & 0.06 & -0.17 & 0.19 \\ -2.85 & -3.80 & -0.69 & -0.07 & 0.19 \\ -1.12 & 0.56 & -0.55 & 0.42 & 0.11 \\ 1.43 & 1.27 & 0.26 & 0.03 & -0.22 \\ 1.62 & 2.28 & 0.82 & -0.28 & 0.20 \end{bmatrix}. \quad (15)$$

Figure 2 shows the first two principal components of the global analysis along with the wine projections for the experts. Note that the position of each wine in the global analysis is the barycenter (i.e., centroid) of its positions for the experts. To facilitate the interpretation, we have drawn lines linking the expert wine projection to the global wine position. This picture shows that Expert 3 is at variance with the other experts, in particular for Wines 3 and 6.

The Original Variables and the Global Analysis

As in standard PCA, the variable loadings are the correlation between the original variables and the global factor scores (cf. Table 2). These loadings are plotted in Figure 3 along with the “circles of correlation.” This figure shows that Expert 3 differs from the other experts and is mostly responsible for the second component of the compromise.

The Original PCAs and the Global Analysis

MFA starts with a series of PCAs. Their relationship with the global analysis is explored by computing loadings (i.e., correlations) between the components of each study and the components of the global analysis. These loadings, given in Table 2, are displayed in Figure 3. They relate the original PCA and the global analysis.

Analyzing the Between-Study Structure

The relationships between the studies and between the studies and the global solution are analyzed by

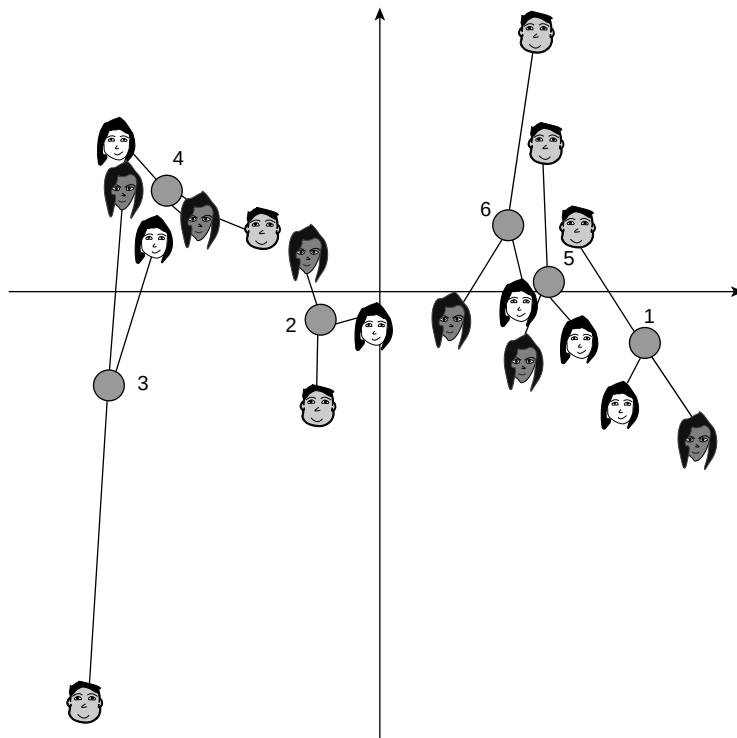


Figure 2 Projection of the Experts Onto the Global Analysis

Notes: Experts are represented by their faces. A line segment links the position of the wine for a given expert to its global position. First component: $\lambda_1 = 2.83$, explains 84% of the inertia; second component: $\lambda_2 = 2.83$, explains 11% of the inertia.

computing the partial inertia of each study for each dimension of the global analysis. This is computed, for each study, as the sum of the squared projections of the variables on the right singular vectors of $\mathbf{Z}$ (cf. Equation 7) multiplied by the corresponding eigenvalue. Because the singular vectors are normalized, the sum of the partial inertias for all the studies for a given dimension is equal to its eigenvalue. For example, for Study 1 and Component 1, the partial inertia is obtained as

$$\lambda_1 \times \sum_j q_{j,1}^2 = 2.83 \times [(-.34)^2 + (.35)^2 + (.32)^2] = 2.83 \times .34 = .96. \quad (16)$$

Similar computations give the values reported in Table 3. These values are used to plot the studies as shown in Figure 4. The plot confirms the originality of Expert 3 and its importance for Dimension 2.

Table 2 Loadings on the Principal Components of the Global Analysis of the Original Variables and the Principal Components of the Study PCAs

Loadings With Original Variables												
Axis	λ	%	Expert 1			Expert 2				Expert 3		
			Fruity	Woody	Coffee	Fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
1	2.83	85	-0.97	0.98	0.92	-0.89	0.96	0.95	0.97	-0.59	0.95	0.99
2	.36	11	0.23	-0.15	-0.06	0.38	-0.00	-0.20	0.10	-0.80	0.19	0.00
3	.12	3	0.02	-0.02	-0.37	-0.21	0.28	-0.00	-0.14	0.08	0.24	-0.11

Loadings With First Two Components From Study PCAs								
Axis	λ	%	Expert 1		Expert 2		Expert 3	
			PC1	PC2	PC1	PC2	PC1	PC2
1	2.83	85	.98	.08	.99	−.16	.94	−.35
2	.36	11	−.15	−.28	−.13	−.76	.35	.94
3	.12	3	−.14	.84	.09	.58	.05	−.01

Note: Only the first three dimensions are kept.

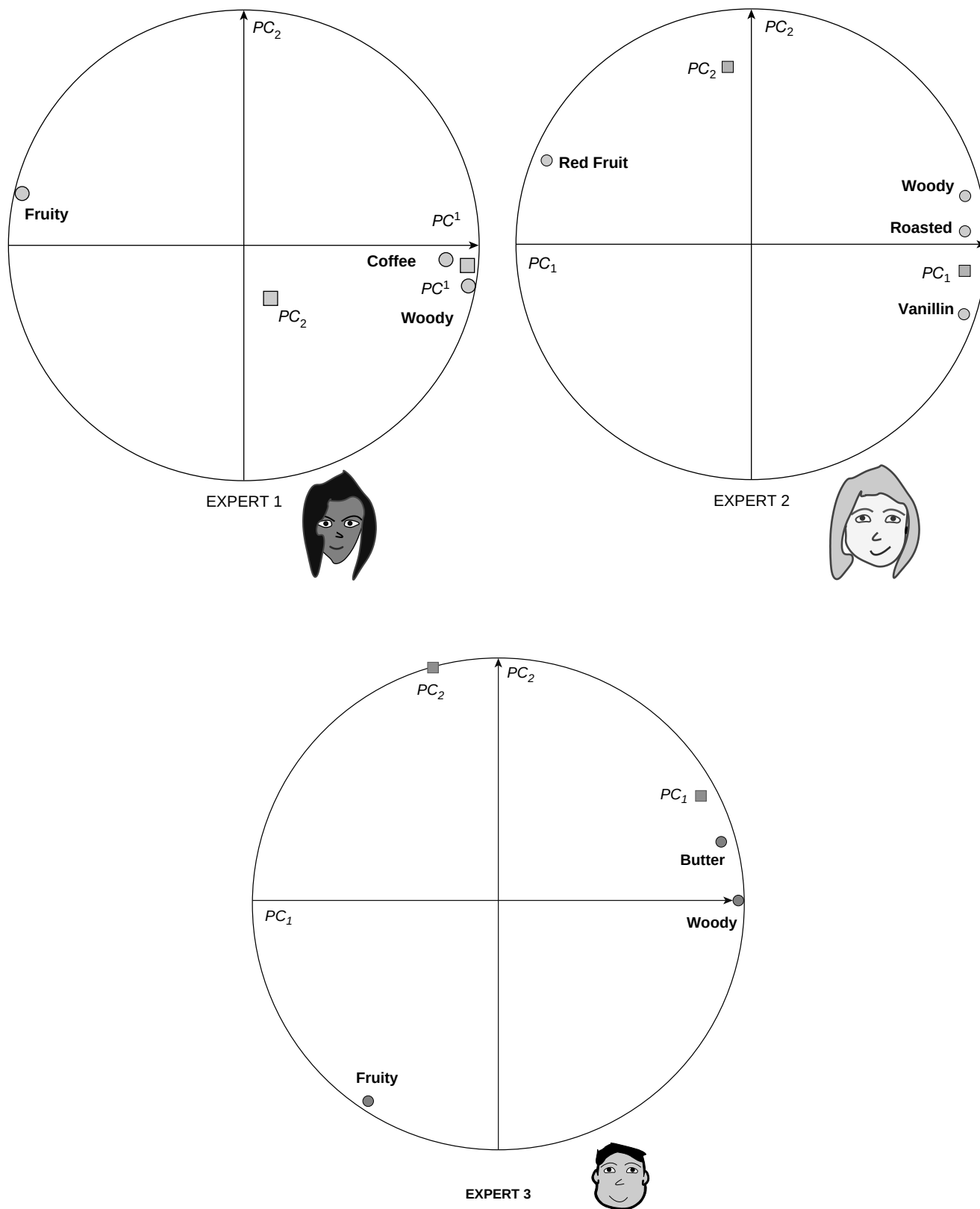
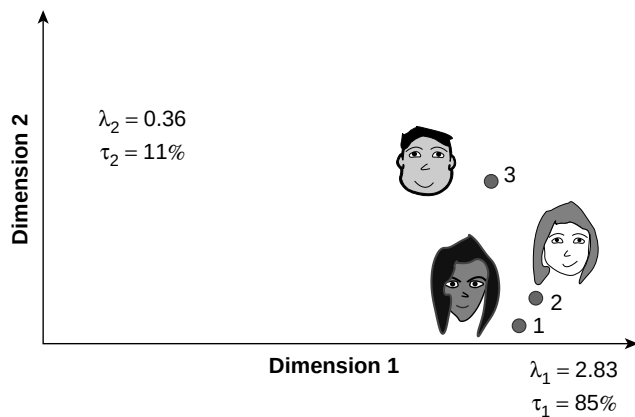


Figure 3 Circle of Correlations for the Original Variables

Table 3 Partial Inertias for the First Three Dimensions

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Expert 1	0.96	0.03	0.05	0.01	0.01
Expert 2	0.98	0.06	0.04	0.02	0.00
Expert 3	0.90	0.28	0.03	0.00	0.00
Σ	2.83	.36	.12	.03	.01
	λ_1	λ_2	λ_3	λ_4	λ_5

**Figure 4** Partial Inertia: Plot of the Experts on the First Two Components

—Hervé Abdi and Dominique Valentin

See also Centroid; DISTATIS; Eigendecomposition; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Principal Component Analysis; R_v and Congruence Coefficients; Singular and Generalized Singular Value Decomposition; STATIS

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Escofier, B., & Pagès, J. (1990). *Analyses factorielles simples et multiples: Objectifs, méthodes, interprétation*. Paris: Dunod.
- Escofier, B., & Pagès, J. (1994). Multiple factor analysis. *Computational Statistics & Data Analysis*, 18, 121–140.

MULTIPLE IMPUTATION FOR MISSING DATA

In many, if not most, studies, some data that were meant to be collected are missing. For example, in a survey, some people may not respond to all the questions. Or, in a randomized experiment, some units' outcomes may not be measured because of equipment failure. Multiple imputation is one principled method for handling such missing data. The general idea is to fill in the missing data with plausible values, analyze the completed data set, and repeat the process multiple times. The analyses from each completed data set are combined to result in inferences that properly account for the missing data. Multiple imputation has been used in large governmental surveys such as the National Health and Nutrition Examination Survey and the Survey of Consumer Finances, and in numerous studies by individual researchers.

Before we review multiple imputation, it is worthwhile to consider the most common and convenient approach to handling missing data: Analyze only the cases that have complete data for the variables of interest. This available cases approach can lead to inaccurate estimates. For a simple illustration of this point, consider the hypothetical data in Table 1 for a random sample of five people. Suppose that weights of all people over 6 feet tall are missing—so that the observed data are 130, 140, and 150—because the height/weight instrument is unable to record information for people over 6 feet tall. Researchers interested in estimating the population average weight are in

Table 1 Hypothetical Data for Illustrating Multiple Imputation

Height (inches)	Weight (pounds)
65	130
68	140
70	150
72	160
75	170

trouble if they use only the three available cases: their sample average is a severe underestimate.

Many times, researchers are interested in relationships among variables, such as regression coefficients. In the hypothetical example, the fitted regression of weight on height obtained using the three available cases results in reasonable (unbiased) estimates of the slope and intercept, because the regression holds for all heights. However, in data sets with many variables and complicated missing data patterns, using only the available cases might exclude a large fraction of the observations, which could dramatically increase the variability of the estimates. Additionally, different specifications of models may use different units for estimation, making theoretical properties of resulting inferences nearly impossible to understand and practical comparisons of different models difficult.

Illustration of Multiple Imputation

In contrast to available cases analyses, multiple imputation uses all records for estimation, which takes advantage of the information from partially completed records. To illustrate multiple imputation, we again use the hypothetical example. We first demonstrate how to analyze a set of multiply imputed data sets, and then discuss methods of generating imputations.

Suppose that five plausible values for each missing weight have been generated to create five completed data sets. These are displayed in Table 2, along with the estimated slope and its variance obtained from fitting standard linear regression in each completed data set. Inferences for the population regression slope β

are based on three quantities. First, compute the average of the five estimated slopes, which equals 4.01. Second, compute the variance of these five estimated slopes, which equals .0523. Third, compute the average of the variances of the slopes, which equals .0209. The point estimate of the population slope is 4.01, and the variance associated with this point estimate is $(1 + 1/5)(0.0523) + 0.0209 = 0.0836$. An approximate 95% confidence interval for β is $4.01 \pm 1.96\sqrt{.0836}$. A similar process can be followed to obtain inferences for the population average weight, using the sample average weights and their variances. This resulting approximate 95% confidence interval is $149.5 \pm 1.96\sqrt{48.63}$.

Generically, suppose the goal is to estimate some parameter Q , such as a population mean or a regression coefficient. Let q be the estimator of Q , and let u be the estimator of the variance of q . In each completed data set l , where $l = 1, \dots, m$, let q_l and u_l be the values of q and u obtained from that data set. The following quantities are required for inferences:

$$\begin{aligned}\bar{q}_m &= \sum_{l=1}^m q_l \\ b_m &= \sum_{l=1}^m (q_l - \bar{q}_m)^2 / (m - 1) \\ \bar{u}_m &= \sum_{l=1}^m u_l.\end{aligned}$$

The point estimate of Q is $\bar{q}_m$, and its variance is estimated by

$$T_m = (1 + 1/m)b_m + \bar{u}_m.$$

An approximate 95% confidence interval is

$$\bar{q}_m \pm 1.96\sqrt{T_m}.$$

The use of a normal distribution for confidence intervals is most appropriate when the number of records in the completed data set is large, say, 40 or more, and the number of completed data sets m is

Table 2 Five Completed Data Sets After Imputations

Height	Weight Completed 1	Weight Completed 2	Weight Completed 3	Weight Completed 4	Weight Completed 5
65	130	130	130	130	130
68	140	140	140	140	140
70	150	150	150	150	150
72	157	166	155	157	156
75	171	169	167	171	168
Estimated					
Slope	4.12	4.26	3.71	4.12	3.83
(Variance)	(0.025)	(0.346)	(0.024)	(0.025)	(0.018)

large, say, 20 or more. When the number of records is large but m is moderate, inferences should be made using a t distribution with degrees of freedom v_m , where

$$v_m = (m - 1) \left(1 + \frac{\bar{u}_m}{(1 + 1/m)b_m} \right)^2.$$

When the number of records and m are moderate, we use a t distribution with a complicated degrees of freedom formula that can be found in the references below. Using these t distributions rather than normal distributions improves the coverage properties of the confidence intervals.

Imputing missing values is a complicated process for complex patterns of missing data. In general, the analyst first specifies a statistical model relating all the variables in the data set. The analyst then uses that model and the observed data to estimate the distributions of the model parameters (e.g., the intercept, slope, and regression variance in the hypothetical data). Based on these estimated distributions, the analyst generates plausible values of the model parameters. Using these values and the observed data, the analyst generates values of the missing data. This creates one completed data set. The process is repeated m times.

Generating imputations from scratch requires knowledge of Bayesian statistics and Markov chain Monte Carlo methods. Fortunately, software exists that makes it unnecessary to generate imputations from scratch.

Software for Multiple Imputation

When data can be reasonably described by a multivariate normal distribution, analysts can use the free software package NORM, written by Joe Schafer, to generate and analyze multiply imputed data sets. This is a standalone package that runs on PCs. The NORM Web site also has links to software that does multiple imputation for purely categorical data (CAT), data that are mixed categorical and multivariate normal (MIX), and longitudinal data (PAN). These packages run within the S Plus software environment.

Another approach is to construct a series of regression models, using each variable with missing values as the outcome variable in one of the regressions. For example, suppose age is measured in the hypothetical

data set, and its value is missing for some people. Missing weights are imputed from a regression of weight on height and age, and missing ages are imputed from a regression of age on height and weight. This approach allows analysts to specify models one variable at a time, which eases computations. This approach is implemented in the free software IVEWARE for SAS, written by Trivellore Raghunathan. It is also implemented in the free software package MICE for S Plus.

Assumptions of Multiple Imputation

Two conditions are needed for inferences from the multiply imputed data sets to be valid, in the sense that 95% of all 95% confidence intervals contain their parameter of interest over repeated application. First, if there were no missing data, the estimator q and its variance u should produce a valid 95% confidence interval for the parameter of interest. For example, the relationship between weight and height needs to be well described by a linear regression if we are to trust the 95% confidence interval for the slope. Second, the imputations need to be plausible values, in the sense that they come from a model that faithfully reflects the relationships in the data. For example, if weights are implausibly imputed by plugging in the average weight rather than using the regression on heights, the relationship between weight and height would be distorted.

The imputation process does not affect the validity of the first condition, but it is central to the second condition. It is essential that imputations be made from expansive models. For example, include as many predictors as possible in the regression models. This ensures that those relationships will be reflected in the imputations.

It is necessary that $m > 1$; otherwise, it is not possible to estimate b_m . Intuitively, if the analyst imputes just one data set and then analyzes it, the analyst essentially would be acting as if the imputed values were genuine values, when in reality, they were simulated from probability models. Generating multiple plausible values allows the uncertainty due to imputation to be passed on to inferences. In general, the larger m is, the more accurate inferences tend to be. Research has shown that $m = 5$ is often adequate, especially when the fraction of missing values is not large.

Extensions to Multiple Imputation

In addition to the inferential methods for scalar estimands presented here, researchers have developed methods for doing significance tests for multi component estimands. For example, it is possible to perform likelihood ratio tests of nested models using multiply imputed data sets. Multiple imputation also has been suggested as a way to protect data confidentiality.

—Jerome Reiter

See also Missing Data Method

Further Reading

- Barnard, J., & Rubin, D. B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Multiple imputation software: <http://www.stat.psu.edu/~jls/misoftwa.html>

MULTITRAIT MULTIMETHOD MATRIX AND CONSTRUCT VALIDITY

A multitrait multimethod matrix (MTMM) refers to a matrix of correlations derived from a factorial design in which each of a set of constructs has been assessed with each of a set of different methods of measurement. Originally developed by Donald T. Campbell and Donald W. Fiske, the MTMM was proposed to assess construct validity (i.e., the degree to which an operationalization reflects its intended construct) using two subcategories of validity. “Convergent validity” is the extent to which measures intended to assess the same construct are associated with one another. “Discriminant validity” is the extent to which measures designed to assess different constructs are, in fact, distinct from one another.

The goal of the MTMM is to provide a framework to assess the effects of trait variance (variance attributed to the intended construct of interest) and

method variance (variance attributable to the specific method of measurement) by examining convergent and discriminant validity. More generally, the MTMM also provides information about patterns of associations between methods, and patterns of associations between constructs and possible interactions between methods and constructs.

Selecting Traits and Methods for an MTMM

The selection of traits and methods for an MTMM analysis should follow a set of guidelines. First, each method should be well suited to evaluating all of the constructs of interest. Second, measurement methods should be as independent of each other as possible. Finally, constructs should be included that are expected to vary in their degree of association. In other words, the matrix should include some constructs postulated to be highly related and some constructs thought to be unrelated. This range of association is recommended in order to provide a basis of comparison for discriminant validity.

Organization of an MTMM

Once MTMM data are collected, correlations among the observed scores are computed and arranged with a specific organization. This can be illustrated with a hypothetical example (see Table 1). This example has three traits (feelings, beliefs, and intentions to act), each of which has been assessed using three measurement methods: a Thurston equal appearing interval scale, a Likert summated ratings scale, and a Semantic Differential scale. Thus, there are nine observations for each person, thereby producing a 9×9 correlation matrix.

The MTMM is organized by method, with each of the three constructs embedded in each method block. The interpretation of the MTMM involves identifying the four different types of construct method pairings within the matrix, which take the form of either a diagonal or a triangle. These pairings are evaluated using a set of guidelines.

The first type of construct method pairing is the monotrait heteromethod diagonals, which are

Table 1 Example of a Multitrait-Multimethod Matrix

Measure	Thurstone			Likert			Semantic Differential		
	<i>F</i>	<i>B</i>	<i>I</i>	<i>F</i>	<i>B</i>	<i>I</i>	<i>F</i>	<i>B</i>	<i>I</i>
Thurstone									
Feeling	(.98)								
Belief	<u>.62</u>	(.95)							
Intentions	<u>.19</u>	<u>.17</u>	(.93)						
Likert									
Feeling	.75	.60	.18	(.95)					
Belief	.59	.86	.17	<u>.60</u>	(.94)				
Intentions	.19	.18	.74	<u>.21</u>	<u>.20</u>	(.95)			
Semantic Differential									
Feeling	.78	.59	.20	.76	.60	.17	(.99)		
Belief	.60	.88	.19	.59	.84	.19	<u>.64</u>	(.95)	
Intentions	.19	.20	.77	.18	.18	.77	<u>.20</u>	<u>.20</u>	(.92)

Note: Monotrait-monomethod values are in parentheses; heterotrait-monomethod values are underlined; monotrait-heteromethod values are in bold; heterotrait-heteromethod values are unmarked.

correlations between different measures of the same construct. These correlations are in bold and include, for example, the correlation between feelings evaluated by both the Likert Scale and the Thurstone scale ($r = .75$). These correlations are also known as the validity diagonals.

The second type of construct method pairing is the heterotrait heteromethod triangles, which are the correlations between different measures of different traits (i.e., correlations between pairings that have neither trait nor measure in common). These correlations are unmarked in the matrix and include, for example, the correlations between feelings evaluated by the Likert scale and beliefs evaluated by the Thurstone scale ($r = .60$).

The third type of construct method pairing is the monotrait monomethod diagonals, which are the correlations between the same constructs using the same measures. These correlations are highlighted in the matrix via parentheses and include feelings evaluated by both measures of the Thurstone scale ($r = .98$). These are essentially the correlation of the measure with itself, and for that reason, these correlations are estimates of the reliability of each measure.

The fourth type of construct method pairing is the heterotrait monomethod triangles, which are the correlations between different constructs evaluated using the same measure. These correlations are underlined in the matrix and include feelings and beliefs evaluated by the Thurstone scale ($r = .62$).

Interpreting the MTMM

The guidelines for interpreting an MTMM begin by evaluating the monotrait monomethod (reliability) diagonal. These coefficients should be consistently the highest in the matrix because it is unlikely that a measure will correlate more highly with anything but itself. Following our example, the monotrait monomethod correlations are indeed the highest of the matrix, ranging from .92 to .99.

Second, the monotrait heteromethod correlations are used to assess convergent validity. Large correlations of this sort indicate convergent validity because the different methods are in agreement on the same underlying constructs. In our example, the monotrait heteromethod correlations are high (ranging from .74 to .88), thereby suggesting that different methods produce similar results for the three constructs.

Although high monotrait heteromethod correlations are a necessary condition for convergent validity, they are not sufficient for its existence. Specifically, even when monotrait heteromethod correlations are large, it is possible that they may be artificially inflated based on an irrelevant factor (e.g., different measures may be highly correlated because of method variance). Thus, the values in the monotrait heteromethod diagonals should be compared to the values in the heterotrait heteromethod triangles. Specifically, the values of the monotrait heteromethod diagonal should be higher than those of the heterotrait heteromethod correlations that share the same column or row. Essentially, if these different methods are truly measuring the same construct (i.e., it is mainly trait variance responsible for the high monotrait heteromethod correlations), these values should be correlated more strongly than unrelated constructs that are measured using different methods (i.e., values having neither construct or method in common). For example, in the sample matrix, the correlation between the Thurstone and Likert scales evaluating feelings ($r = .75$) is higher than the correlation between the Likert scale evaluating feelings with the Thurstone scale evaluating behavior ($r = .60$), the Likert scale evaluating feelings with the Thurstone scale evaluating intentions ($r = .18$), and so on. A visual examination reveals that the monotrait heteromethod values are higher than the heterotrait heteromethod values for each comparison in the matrix. Furthermore, heterotrait heteromethod correlations for constructs hypothesized to be related should be higher than for those not hypothesized to be related.

Next, monotrait heteromethod values should be compared to the heterotrait monomethod triangles. Specifically, the values in the monotrait heteromethod diagonal should be higher than those of heterotrait monomethod triangles, because different methods measuring the same trait should be more highly correlated than the same methods measuring different traits. If the heterotrait monomethod correlations are larger, it suggests that the method of measurement may account for a larger proportion of the observed variance in the scores. In our example,

the correlation between the Thurstone and Likert evaluations of feelings ($r = .75$) is higher than the Thurstone values for feelings and belief ($r = .62$), the Thurstone values for feelings and intentions ($r = .19$), and the Thurstone values for beliefs and intentions ($r = .17$). This criterion is met for each comparison in the matrix.

Finally, researchers should evaluate both the heterotrait monomethod and the heterotrait heteromethod triangles. If two traits are correlated, this relationship should hold irrespective of the method used to evaluate them, and therefore, the same pattern of interrelationships between different traits should be visible in all of the monomethod and heteromethod blocks. In examining the sample matrix, this criterion is satisfied (e.g., feelings and beliefs are always more highly correlated than are feelings and intentions).

Limitations of the MTMM

Although the MTMM remains a popular and useful approach, it does have limitations. For example, many key underlying assumptions are not clearly defined (e.g., the MTMM assumes the existence of both method and trait factors but does not specify their relationships with each other). It also provides no manner of testing its underlying assumptions and is unrealistic in that it neglects to account for random measurement error. Additionally, there are practical problems associated with the use of MTMM. For example, complete cross factoring of methods and constructs is not always possible, and very large matrices can be difficult to evaluate. The MTMM approach has also been criticized for its ambiguity. MTMM matrices sometimes produce conflicting results within a given matrix. Thus, in practice, some aspects of the matrix may be consistent with the guidelines, whereas others may not. In such cases, evaluations of the matrix can be very subjective.

In order to address the difficulty of quantifying the extent to which MTMM criteria have been satisfied, statistical procedures have been proposed to complement the visual evaluations of the matrix. The most common approach is the use of confirmatory factor

analysis. A number of different factor analytic models have been proposed to evaluate MTMM data. Unfortunately, many of these models have conceptual and practical limitations. Research evaluating the strengths and weaknesses of these models remains an active area of inquiry in the quantitative methods literature, and some promising techniques have been identified (e.g., the Composite Direct Product Model). However, definitive comparisons of many of these models have yet to be conducted.

— *Leandre R. Fabrigar
and Marie Joelle Estrada*

See also Construct Validity; Validity Theory

Further Reading

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112, 393–395.
- Pardy, S. A., Fabrigar, L. R., & Visser, P. S. (2005). Multitrait multimethod analyses. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in the behavioral sciences* (pp. 1343–1347). Chichester, UK: Wiley.
- Shrout, P. E., & Fiske, S. T. (Eds.). (1995). *Advances in personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Hillsdale, NJ: Erlbaum.

Multitrait-Multimethod Matrix: <http://www.socialresearchmethods.net/kb/mtmmmat.htm>

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Multivariate analysis of variance, or MANOVA, is a data-modeling technique that is a powerful alternative to its univariate analysis of variance (ANOVA) counterpart. With traditional ANOVA, mean differences between groups on a single quantitative variable can be analyzed. For instance, consider a school psychologist who is interested in examining the average performance of Asian, Caucasian, and Hispanic children on a standardized test of intelligence that yields

continuous scores on three subscales: Mathematics, Verbal, and Processing Speed. Adopting a univariate frame of mind, the psychologist could conduct three separate ANOVAs to determine if the means of the ethnic groups are equivalent on each of the three subscales. If the psychologist viewed the data from a multivariate standpoint, however, he or she could conduct a single MANOVA to examine differences between the Asian, Caucasian, and Hispanic children on linear combinations of the three intelligence subscales. In other words, he or she could combine the Mathematics, Verbal, and Processing Speed scores to form one or more multivariate composites on which the three ethnic groups could be compared. Perhaps a combination of high Mathematics, high Processing Speed, and low Verbal scores will provide the most effective composite for discriminating among the three groups? Perhaps the simple difference between the Mathematics and Verbal subscale scores will instead provide the most effective composite? Such questions are multivariate in nature because they address the interrelationships among the continuous measures of intellectual ability.

The primary strength of MANOVA is thus the capability to examine group differences on linear combinations of quantitative variables. For the sake of convenience, grouping variables will herein be referred to as *independent variables*, and the quantitative variables will be referred to as *dependent variables*. In the example above, ethnicity (with three levels: Asian, Caucasian, and Hispanic) is the independent variable, and the three intelligence subscales are the dependent variables. Mathematically speaking, there are no limits to the number of independent variables or dependent variables that can be included in the analysis, and the independent variables can be composed of two or more groups. When only one independent variable is included, the analysis is referred to as a one-way (or one-factor) MANOVA, and when two or more independent variables are included, the analysis is referred to as a factorial MANOVA. In the special case of one independent variable with two levels, the analysis is often referred to as a Hotelling's *T*, which is the multivariate generalization of the independent samples *t* test. Regardless

of the study design, the goal of the MANOVA is to determine if the independent variable groups differ in their means on at least one linear combination of the dependent variables.

An Example

Consider a study of the *Sesame Street* television series reported by James P. Stevens. Children in the 3–5 year age range were studied and assessed on a variety of measures both before and after watching episodes of *Sesame Street*. One question that could be addressed in this data regards differences between boys and girls prior to viewing the episodes. Measures of knowledge in four areas are of particular interest: body parts, letters, forms, and numbers. Scale values on these measures range from 0–58, 0–20, 0–54, and 0–32, respectively. Are the mean performances of the boys and girls similar on these scales? Gender, with two levels (boys and girls), will be treated as the independent variable, and the four knowledge scales will serve as the dependent variables.

For pedagogical reasons only, the data are first analyzed with four ANOVAs, examining each knowledge area separately. The null hypothesis for each analysis represents equality between population means on the variable; $H_0: \mu_{\text{boys}} = \mu_{\text{girls}}$, where μ represents the population mean. The resulting F values and associated statistics computed from SPSS are shown in Table 1. Using an unadjusted a priori p value of .05, it can be seen that the ANOVAs are not statistically significant (all observed, “Sig.,” observed p values > .05). The results therefore suggest that the population means for the boys and girls are equal in each of the four knowledge areas.

These data are also suitable for a one way MANOVA given the single independent variable and multiple dependent variables. The null hypothesis for this analysis states that the two groups

Table 1 Univariate ANOVA Results

Variable		Sum of Squares	df	Mean Square	F	Sig.
Body Parts	Gender Effect	81.809	1	81.809	2.011	.157
	Residual	9679.791	238	40.671		
	Total	9761.600	239	40.844		
Letters	Gender Effect	.639	1	.639	.009	.926
	Residual	17415.423	238	73.174		
	Total	17416.063	239	72.871		
Forms	Gender Effect	16.153	1	16.15	1.157	.283
	Residual	3321.343	238	13.955		
	Total	3337.496	239	13.964		
Numbers	Gender Effect	10.417	1	10.417	.091	.763
	Residual	27277.978	238	114.613		
	Total	27288.396	239	114.177		

are equal on a linear combination of all four dependent variables— $H_0: \mu_{\text{linear,boys}} = \mu_{\text{linear,girls}}$. The mathematics underlying the MANOVA will produce the linear combination in such a way as to maximize the mean difference between the boys and girls. The mean difference on the linear combination then can be tested for statistical significance, as shown in Table 2.

Four tests of statistical significance are commonly available for MANOVA, and three of these tests have well known sampling distributions from which p values can be estimated. The significance level of the remaining test, Roy’s *greatest characteristic root* (or g.c.r., with s , m , and n degrees of freedom from Table 2), can also be estimated with special software developed by Richard J. Harris.

In this example, the tests show that the results of the MANOVA are statistically significant; in other words, the population means for the boys and girls

Table 2 MANOVA Results

Multivariate Tests of Significance ($s = 1$, $m = 1$, $n = 116 \frac{1}{2}$)					
Test Name	Value	Exact F	Hyp. df	Error df	Sig. F
Pillais	.04493	2.76403	4.00	235.00	.028
Hotelling	.04705	2.76403	4.00	235.00	.028
Wilks	.95507	2.76403	4.00	235.00	.028
Roy	.04493				

are judged to be different on a linear combination of the four areas of knowledge. The F values and p values are all identical because only one independent variable with two levels is included in the analysis. In other study designs, these values will not necessarily be equal. Furthermore, in this example, only one linear combination of the dependent variables (to be shown below) is being tested. With more complex designs, two or more linear combinations will be simultaneously computed and included in these statistical tests, and the four multivariate tests above will not be equivalent. In such instances, one of the four tests must be used, and Wilks's Lambda is the most popular choice. Roy's g.c.r., however, is arguably most suitable for conducting follow-up statistical tests on the linear combinations of dependent variables. The readings cited below provide explanations of the differences among the tests.

The next crucial step in the analysis is to determine how the dependent variables are combined to significantly discriminate between the boys and girls. A number of methods are available for determining the importance or relevance of each dependent variable to the multivariate effect. The approach adopted here relies on the *discriminant function coefficients* produced from the MANOVA. These coefficients are regression weights derived to maximally discriminate between the groups that comprise the independent variable. The coefficients can be computed in both raw and standardized form, and the values for the current example are reported in Table 3.

If the dependent variables are measured on different scales, as in this example, the standardized coefficients should be examined; otherwise, the raw

coefficients can be used. In either case, the absolute relative magnitudes of the coefficients are examined, and the current results indicate that the Forms and Body Parts dependent variables are contributing most to the multivariate effect. The signs of the discriminant function coefficients for the important variables are next considered to interpret the linear combination. In this example, the multivariate composite clearly conveys a discrepancy effect; specifically, the boys and girls differ in their mean discrepancy between the Forms and Body Parts variables. The nature of this effect can be understood by examining the means of the standardized dependent variables for both boys and girls in Figure 1.

As can be clearly seen, compared to the girls, the boys performed better on the Forms knowledge scale and worse on the Body Parts knowledge scale. This effect is also reflected in the means for the boys ($M = -.22$, $SD = .85$) and girls ($M = .21$, $SD = 1.12$) on the standardized multivariate composite (i.e., the standardized linear combination). The lower mean value for the boys indicates superior performance on the Forms scale relative to the Body Parts scale. This intriguing multivariate effect was completely missed by the univariate ANOVAs reported above!

Additional Issues

The example above includes four dependent variables and only a single independent variable with two groups. It is therefore among the simplest of designs for a MANOVA, and it yielded only a single multivariate composite (linear combination). For more complex designs, additional composites will be produced from the analysis depending on the number of groups in the independent variables, the number of independent variables, and the number of dependent variables. The multivariate tests, except Roy's g.c.r., will then address a null hypothesis that posits equality among population group means on all of the multivariate composites, considered simultaneously. Roy's g.c.r. addresses the null hypothesis that the groups differ in the population means on only the first multivariate composite. When multiple composites are produced, the first will maximally discriminate

Table 3 Discriminant Function Coefficients

Variable	Raw	Standard
Body Parts	.22671	1.44580
Letters	.03172	.27134
Forms	-.28492	-1.06436
Numbers	-.05452	-.58372

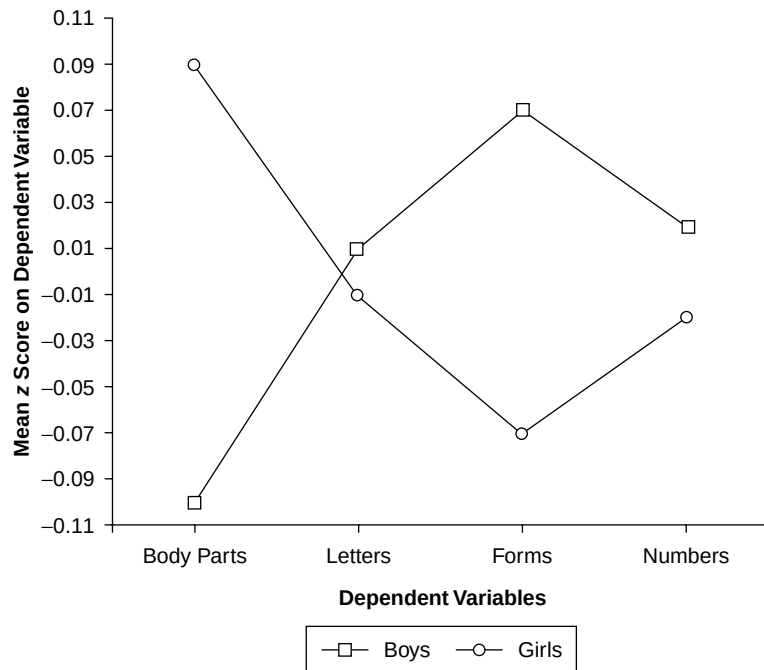


Figure 1 Profile Plot for Boys and Girls on Four Dependent Variables

among the groups, and the remaining composites will be ordered according to the magnitudes of group differences they yield. The composites will all be orthogonal (or uncorrelated), however, and thus can be examined and interpreted separately, using the techniques described above. To make matters even more complex, factorial MANOVAs yield tests of interactions and main effects. For each interaction and main effect, a unique set of multivariate composites will be generated that must be evaluated. With complex study designs, it is easy to see that performing a MANOVA is no simple task.

Regardless of study design, it should be stressed that the heart of MANOVA is the multivariate composite or set of composites generated from the analysis. Unfortunately, many investigators fail to examine the linear combinations of the dependent variables. Instead, two misuses of the analysis are common. First, univariate ANOVAs are conducted after a MANOVA to determine the importance of the dependent variables in the multivariate effect. This strategy is gravely erroneous because, as was clearly demonstrated in the example above, the multivariate effect conveyed in the discriminant

function coefficients will not necessarily correspond to the univariate effects found in the separate ANOVAs. Second, MANOVA is often used as a method for controlling Type I error inflation. Many investigators will first conduct a MANOVA, and if the result is statistically significant, they will then conduct separate ANOVAs and judge the results using an unadjusted *a priori* *p* value of, for example, .05. Unfortunately, conducting a MANOVA alone will not normally provide protection against Type I error inflation, and this analysis strategy is not to be condoned.

Finally, MANOVA is a statistical modeling technique and thus involves inferential judgments about population parameters. As with all statistical tests, the validity of such inferential leaps

relies on a number of assumptions. In MANOVA, these are (a) independence of observations, (b) the dependent measures follow a multivariate normal distribution for each level of the independent variable, and (c) the population covariance matrices for the dependent variables are equal for all levels of the independent variable. Assessing each of these assumptions is an important part of any MANOVA.

—James W. Grice

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA)

Further Reading

- Harris, M. B., Harris, R. J., & Bochner, S. (1982). Fat, four eyed, and female: Stereotypes of obesity, glasses, and gender. *Journal of Applied Social Psychology*, 12, 503–516.
- Harris, R. J. (1993). Multivariate analysis of variance. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 255–296). New York: Marcel Dekker.
- Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Erlbaum.
- Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and*

mathematical modeling (pp. 183–208). San Diego, CA: Academic Press.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

MULTIVARIATE NORMAL DISTRIBUTION

In 1984, Anderson stated, “A major reason for basing statistical analysis on the normal distribution is that this probabilistic model approximates well the distribution of continuous measurements in many sampled populations.” He goes on to mention that normality based methods have the advantage that the theory is developed in great detail. In univariate analysis, the central limit theorem steers us toward the normal distribution. Similarly, we are steered toward the multivariate normal distribution in multivariate analysis by the general, or multivariate, central limit theorem.

As its name would indicate, the multivariate normal distribution is the multidimensional extension of the familiar univariate normal distribution, or “bell curve.” The bivariate normal distribution will arise when a pair of variables X and Y has not just individual (or marginal) distributions that are normal, but also a joint distribution that is normal. The bivariate normal distribution can be visualized as a three-dimensional bell. The distribution can be extended

with three or more variables sharing this relationship.

Properties

The probability density function for a univariate normal distribution is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)/\sigma]^2/2}$, where $-\infty < x < \infty$; μ represents the mean of the distribution and can take on any real number value; and σ represents the standard deviation of the distribution, which can take on any positive value. This formula is extended to the multivariate distributions as $f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{0.5}} e^{-[(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)]/2}$. Many online sites, including Weisstein’s MathWorld, cover many of the mathematical properties of this distribution.

For the bivariate distribution, $p = 2$, and in general, p is the number of variables. μ is a vector of size $p \times 1$ that contains the means, and Σ is the variance covariance matrix of size $p \times p$. In this matrix, the main diagonal elements are the variances of the p component variables, and the off diagonal elements are the covariances. For example, if we let σ_{ij} represent the element in the i th row and j th column of this matrix, then σ_{11} would represent the variance of the first variable, σ_{12} is the covariance between the first and second variables, and the correlation between these two variables would be $r_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$, or the covariance divided by the product of the two standard deviations.

Some properties of the multivariate normal distribution are very important to consider when performing multivariate analyses. When considering a random vector $\mathbf{X}$ that follows a multivariate normal distribution, the following properties will hold:

1. All linear combinations of the components of $\mathbf{X}$ are normally distributed.
2. All possible subsets of the components of $\mathbf{X}$ will be normally distributed.
3. Zero covariance between two components implies that those components are independent.
4. The conditional distributions of the components are normal.

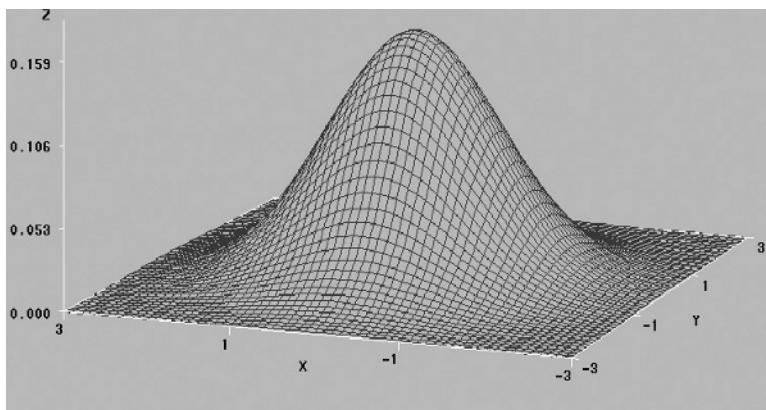


Figure 1 Example of a Multivariate Normal Distribution

It is important to realize that even though each univariate component of $\mathbf{X}$ follows the univariate normal distribution, the univariate normality of each component of some other random vector $\mathbf{X}^*$ does not necessarily imply that $\mathbf{X}^*$ is multivariate normal. In other words, the univariate normality of each variable is necessary but not sufficient to establish multivariate normality. Thus, a strategy of assessing multivariate normality that merely assesses each component variable separately will not be successful in detecting all deviations from multivariate normality. The relationship between the variables, numerically expressed in the variance covariance matrix, must be considered for the problem of testing multivariate normality.

Implications for Analysis

Looney stated that “the assumption of multivariate normality underlies many commonly used multivariate statistical procedures.” Examples include multivariate analysis of variance (MANOVA), discriminant analysis, and canonical correlation. Looney commented that several Monte Carlo studies found that the performance of these multivariate procedures can be adversely affected by a violation of the assumption of multivariate normality.

It is evident that the question of testing for multivariate normality is significant for several reasons. First, ignoring the assumption of multivariate normality when conducting multivariate analyses that call upon that assumption can lead to either increased Type I error rate or a reduction in power. Neither of these possibilities is desirable; in fact, we wish to avoid Type I error and also highly favor statistical methodology that will maximize power. Testing the assumption of multivariate normality will help avoid these undesirable outcomes.

Testing for Goodness of Fit

Many tests exist that will test for multivariate normality. Henze and Mecklin and Mundfrom have provided recent reviews summarizing the dozens of tests that exist. Henze focused primarily on a certain class of

tests that is “affine invariant” and is mathematically consistent against all possible alternatives (i.e., possible deviations from multivariate normality). Although this property of consistency stressed by Henze is mathematically desirable, Mecklin and Mundfrom’s paper also considered many other procedures as well. They felt that other factors besides theoretical properties such as consistency against alternatives should be considered. Mecklin and Mundfrom tended to favor tests that had strong performances in comparison to other candidates in Monte Carlo simulation studies, had test statistics with good asymptotic null distributions and did not require specialized tables, and tests simple enough that those who are not specialists in goodness of fit could perform them. Mecklin and Mundfrom would also value test statistics such as Mardia’s measure of skewness and kurtosis (i.e., multivariate analogues of univariate skewness and kurtosis) that can also serve as multivariate descriptive measures over procedures whose results do not lead themselves well to describing the data.

Many Monte Carlo studies have looked to compare the relative effectiveness of these procedures. Mecklin and Mundfrom compared the Type I and Type II error rates of 13 of these procedures against 25 alternative distributions. As others in the past had found, no single test of goodness of fit to the multivariate distribution is the most powerful in all circumstances. The Henze Zirkler statistic was identified as a test that has both strong theoretical properties and strong performances in Monte Carlo studies. Because the value of the Henze Zirkler test statistic does little to help the researcher diagnose the reason for the deviation from normality, Mecklin and Mundfrom suggested that researchers complement the Henze Zirkler test with the chi square plot described by Healy, and also Mardia’s measures of skewness and kurtosis. Mardia’s measures, particularly kurtosis, are frequently used by structural equation modelers.

Despite the plethora of methods for testing multivariate normality and several power studies assessing their performance, many users of multivariate analyses do not use these tests or seriously consider the assumption of multivariate normality that is made by many multivariate methods. Looney gave five reasons

why he felt that many practitioners do not test for multivariate normality:

1. The practitioner does not know of the existence of tests for multivariate normality.
2. Software for easily calculating the test statistic or p value for a test of multivariate normality is not readily available.
3. Even if software is used to calculate the test statistic, often the significance can be determined only by referencing a special table, which may not be readily available.
4. The practitioner does not want to use a procedure (such as a test of multivariate normality) when little is known about the statistical power of the test.
5. The practitioner is reluctant to test for multivariate normality because he or she does not know how to proceed if non normality is detected.

One reason that an investigator doing applied research in a field outside of statistics might be interested in conducting a test of normality upon a multivariate data set would be to determine the validity of the assumption of normality. Determining whether or not multivariate normality is a tenable assumption is important for deciding on how to proceed in the data analysis. If the assumption is tenable, then the investigator can feel confident in employing the usual parametric techniques of multivariate analysis. However, if the assumption is not tenable, alternatives to parametric statistics need to be considered. Some available options include robust methods, nonparametric procedures, transformations of the data, or computer intensive techniques such as bootstrapping.

—Christopher J. Mecklin

Further Reading

- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.
- Healy, M. J. R. (1968). Multivariate normal plotting. *Applied Statistics*, 17, 157–161.
- Henze, N. (2002). Invariant tests for multivariate normality: A critical review. *Statistical Papers*, 43, 467–503.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *American Statistician*, 49, 64–70.

Mecklin, C. J., & Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72, 123–138.

Mecklin, C. J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75, 93–108.

Rencher, A. C. (1995). *Methods of multivariate analysis*. New York: Wiley.

Weisstein, E. W. (n.d.). *Multivariate normal distribution*. Retrieved from <http://mathworld.wolfram.com/MultivariateNormalDistribution.html>

MYERS-BRIGGS TYPE INDICATOR

The Myers-Briggs Type Indicator (MBTI), a self-report, forced-choice questionnaire, assesses personality according to typologies proposed by Carl Jung. Administration of the MBTI takes approximately 10 minutes. The MBTI is used worldwide in diverse settings; most commonly, it assists in career development, professional team building, personnel selection, and counseling.

Isabel Myers and Katharine Briggs authored the first version of the MBTI in 1962 to measure individual preferences between Jung's personality types. The test, most recently updated in 2001, identifies 16 types. Each is designated by its four-letter code. Each letter of the code indicates a preference for one manner of experiencing the world over another. For instance, an INTJ shows preferences for Introversion (over Extraversion), Intuition (over Sensing), Thinking (over Feeling), and Judging (over Perceiving).

The Introversion Extraversion dimension characterizes how people direct their energy. Introverts direct it internally, toward ideas and the subjective, and extraverts direct it externally, toward people and materials. The Intuition Sensing dimension identifies how people focus their perception. Intuitive types focus readily on abstraction and pattern, whereas

Sensing types focus on concrete stimuli perceived by the senses. Individuals' scores on the Thinking Feeling dimension indicate the types of information on which they base decisions. Thinking types base decisions on attempted rationality and objectivity, whereas Feeling types prioritize maintenance of harmony with others and with their own values. The last dimension, Judging Perceiving, characterizes people's preferences for resolution, among Judging types, or open-ended situations, among Perceiving types. In the most recent revision, each dimension has been parsed into five facets for more nuanced interpretation.

These dimensions allow organizations, for instance, to use the MBTI for team building. After administering the test to a group, members' results can be used to stimulate conversation about their similarities or differences in work style and preferred level of socializing in completing tasks, among other topics. The MBTI provides a common vocabulary with which members can introduce themselves.

Scores are calculated for each dimension based on forced choices between the two sides of the dimension; for example, each item that loads on the Introversion Extraversion scale offers one choice that is weighted in favor of Introversion, and another weighted toward Extraversion. The sides of the four

dimensions for which people most frequently show a preference comprise their type codes. Data used to select items and to provide evidence of validity for MBTI scores were collected from a nationally representative sample. The psychometric qualities for the scales and type codes are somewhat equivocal. Indicators of consistency over time for both scales and type codes and item to item on scales show poorer consistency than other measures frequently used to similar ends.

—Matthew E. Kaler

See also Minnesota Multiphasic Personality Inventory; NEO Personality Inventory; Personality Tests

Further Reading

- McCaulley, M. H. (2000). Myers-Briggs Type Indicator: A bridge between counseling and consulting. *Consulting Psychology Journal: Practice and Research*, 52, 117–132.
- Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Myers & Briggs Foundation: <http://www.myersbriggs.org>

N

A fact is a simple statement that everyone believes. It is innocent, unless found guilty. A hypothesis is a novel suggestion that no one wants to believe. It is guilty, until found effective.

—Edward Teller

NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

According to the National Council on Measurement in Education (NCME) bylaws, “NCME is an organization that is incorporated exclusively for scientific, educational, literary, and charitable purposes.” One primary purpose of NCME is the encouragement of scholarly efforts (a) to “advance the science of measurement in the field of education”; (b) to “improve measurement instruments and procedures for their administration, scoring, interpretation, and use”; and (c) to “improve applications of measurement in assessment of individuals and evaluation of educational programs.”

In addition, NCME has a second primary purpose, the dissemination of information about assessment theory and practices. According to its Web site, NCME disseminates information about (a) “theory, techniques, and instrumentation available for the measurement of educationally relevant human, institutional, and social characteristics”; (b) “procedures appropriate to the interpretation and use of such techniques and instruments”; and (c) “applications of educational measurement in individual and group evaluation studies.”

NCME is a professional organization for persons who are involved in assessment, evaluation, and related aspects of the assessment process. These persons might be involved in test construction; test evaluation; or the generation of new theories and methods for constructing and evaluating tests and test scores, equating test scores, and managing testing at computers, among other things. The membership is diverse, including university faculty, test developers, state and federal testing and research directors, persons working for credentialing organizations, graduate students in measurement and assessment programs, and others who work in the testing field. About 10% to 15% of the 2,000 or so members are from outside the United States, most of them from Canada.

Formation

NCME was established in 1937 and was called the National Association of Teachers of Education Measurement. In 1942, the name was changed to National Council on Measurements Used in Education. The change to the current name came in 1969, along with incorporation.

Publications

NCME publishes through five organs:

1. *Journal of Educational Measurement*
 - Quarterly
 - Publishes technical and theoretical developments in measurement and improvements in the applications of measurement methods
2. *Educational Measurement: Issues and Practice*
 - Quarterly
 - Promotes better understanding of educational measurement and focuses on timely issues and debates in the educational measurement field
3. *Instructional Topics in Educational Measurement Series*
 - Occasional
 - Provides instruction on timely educational measurement topics for practitioners and graduate students
4. Newsletter
 - Quarterly
 - Keeps the membership up to date on current events in the educational measurement field and disseminates information about the activities of the NCME Board and its committees
5. Web site
 - Contains information about the organization, membership, history, awards, publications, and more

Annual Convention

In March or April the organization holds an annual 3-day meeting in conjunction with the American Educational Research Association annual meeting. A high percentage of the membership attends, and one of the highlights is the NCME Breakfast, with attendance of 300 to 400 members. The annual meeting also includes a brief business meeting, awards, and a presidential address.

Achievement Awards

The number of awards given by NCME each year has been on the increase. Currently the following awards are given: Brenda H. Loyd Dissertation Award; Brad Hanson Award; Jason Millman Promising

Measurement Scholar Award; Award for Career Contributions to Educational Measurement; and awards for such things as Outstanding Dissemination of Educational Measurement Concepts to the Public, Outstanding Example of an Application of Educational Measurement Technology, and Technical Contribution to a Field of Educational Measurement. More details about these awards are available at the NCME Web site.

—Ronald K. Hambleton

See also Measurement

Further Reading

Lehmann, I. J. (1990). The state of NCME: Remembering the past, looking to the future. *Educational Measurement: Issues and Practice*, Spring, 3–10.

National Council on Measurement in Education Web site: www.ncme.org

NATIONAL SCIENCE FOUNDATION

The National Science Foundation (NSF) is an independent federal agency created by the U.S. Congress in 1950 to promote scientific progress; advance the health, prosperity, and welfare of the nation; and secure national defense. The NSF's motto, "where discoveries begin," reflects the goal of the foundation, which is to support the people, ideas, and tools that make new knowledge possible. The foundation supports all fields of fundamental science and engineering except the medical sciences. With a current annual budget of \$5.5 billion, the NSF funds one fifth of all federally supported basic research conducted in U.S. colleges and universities and supports more than 200,000 scientists each year. The NSF issues about 10,000 new limited-term (typically 3-year) grants annually, as well as funding for research centers, other research facilities, and scientific equipment.

Organizational Structure

Management at the NSF has two major components: the director, who oversees the staff responsible for

the creation and administration of programs, merit review, and daily operations of the foundation, and the 24-member National Science Board, which establishes the policies of the NSF. The director and all board members serve 6-year terms; they are appointed by the President of the United States and confirmed by the U.S. Senate. The foundation currently employs 1,700 individuals at its headquarters in Arlington, Virginia.

The NSF comprises seven directorates: Biological Sciences; Computer and Information Science and Engineering; Education and Human Resources; Engineering; Geosciences; Mathematics and Physical Sciences; and Social, Behavioral and Economic Sciences. Each directorate has its own assistant director, who oversees its subdivisions. Special divisions, such as the Office of International Science and Engineering, fall under the director's portfolio.

Grants and Merit Review

Grant proposals are submitted by individual scientists or research teams and then assessed through a rigorous merit review system. The NSF offers a number of annual workshops and conferences to discuss scientific discoveries and explore areas requiring new research. The foundation publishes notices (or "solicitations") for new funding opportunities each year and also encourages scholars to submit unsolicited requests for funding in any existing or emerging field of research. The NSF currently receives 40,000 proposals per year, all of which are reviewed by panels of experts not employed by the NSF or by an institution employing a scientist whose proposal is under review. Reviewers are selected from a national pool of experts in each field and asked to provide confidential reviews of proposals.

A Legacy of Discovery

In 2000, in honor of its 50th anniversary, the NSF published *America's Investment in the Future*, providing a history of the organization and its support of science and innovation, including details about more than 100 NSF-funded Nobel Prize winners. The foundation

publishes annual reports on its Web site, as well as the "Nifty 50," a list of NSF-funded inventions, innovations, and discoveries, including the development of the American Sign Language dictionary, nanotechnology, and speech recognition technology.

—Lisa M. Given

Further Reading

National Science Foundation Web site: www.nsf.gov

NEO PERSONALITY INVENTORY

The NEO Personality Inventory–Revised (NEO PI-R; published by Psychological Assessment Resources) is a revision of the NEO PI first published by Paul Costa and Robert McCrae in 1985. The instrument was constructed to measure the "Big Five" personality domains: neuroticism, extroversion, openness, conscientiousness, and agreeableness. In addition to the global measures of these domains, the NEO PI-R includes 30 facet scales, which are subscales related to each of the five major domains.

Two versions of the test booklet are available: a self-report version (Form S) and an observer-report version (Form R). The item content of the two versions is identical, but questions are worded in the first person in Form S and in the third person in Form R. The options of hand scoring and machine scoring are available for both forms, and a computer-based version of the NEO PI-R, which includes administration, scoring, and interpretation, is also available. The most recent manual has been translated into 12 languages other than English.

The NEO PI-R is recommended for use with persons 17 years old and older who do not have psychological disorders that would impair completion of a self-reported measure. Recent research, which included the rewording of some items to increase understanding by adolescents, suggested the NEO PI-R also may be useful with students in high school. To interpret scores, an extensive norm group that includes men and women, including those of

college age, has been collected. The college-age sample is recommended for use with test takers between the ages of 17 and 20 because of elevated scores on the excitement-seeking scale for this age group.

Evidence of validity and reliability has been collected for the NEO PI-R. Extensive research about the structure of the factors (construct validity), as well as correlational data with other measures of personality (convergent validity) and theoretically unrelated constructs (discriminant validity), has been published.

Short-term test-retest reliability coefficients are around .80 for each of the five domains. A 6-year test-retest study found stability coefficients of .68 to .83 for neuroticism, extroversion, and openness. Another study, spanning 7 years, found stability coefficients ranging from .63 to .81 for all five domains. Interitem reliability alpha coefficients for the facet scales range from .86 to .92 for Form S and from .89 to .95 for Form R.

The NEO PI-R has many uses for clients, counselors, and researchers, including increasing clients' self-awareness and helping counselors develop appropriate treatment plans. Additionally, results from a self-report that diverge from a significant other's ratings can give clients insight into the way in which others see them. Coupled with vocational assessments, the NEO PI-R can help clients and counselors discuss possible career decisions on the basis of clients' personality and interests. Furthermore, researchers can benefit by including a measure of personality in their studies.

—Melanie Leuty and Jo-Ida C. Hansen

Further Reading

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor-Inventory Professional Manual*. Odessa, FL: Psychological Assessment Resources.

NEONATAL BEHAVIORAL ASSESSMENT SCALE

The Neonatal Behavioral Assessment Scale (NBAS, published by MacKeith Press) assesses the full range

of neonatal behavior by describing infants' competencies and identifying potential areas of difficulty. The scale was originally developed in 1973 by T. Berry Brazelton on the premise that rather than being passive recipients of their environments, infants actively contribute to the parent-infant relationship. In addition, the scale was designed not only to identify abnormal infant behavior but also to describe the full range of normal infant behavior—a departure from most assessments of neonatal behavior at the time, which focused on identifying developmental delays.

The most recent edition of the NBAS scale takes 20 to 30 minutes to administer and is used to assess full-term infants up to the end of the second month of life. The scale may be adapted to assess preterm and at-risk infants as well. The NBAS is administered by a trained individual, who scores 53 items based on infants' behavior and state. These items assess infants' level of functioning in seven primary domains: (a) infants' reflexes, including the sucking reflex, the plantar reflex, and the rooting response, are assessed; (b) the motor system is assessed by, for example, examining the range of motion and the resistance of infants' limbs; (c) the autonomic system is assessed by examining the degree to which infants are startled during the procedure, how much they tremble during the procedure, and the color of their skin; (d) the capacity of infants to habituate to disturbing stimuli is assessed by shaking a bell nearby or shining a light in the eyes and examining how long it takes infants to disregard the stimulus; (e) infants' state organization is assessed by examining how irritable and excited they are during the procedure, as well as how much their state changes during the procedure; (f) the degree to which infants are capable of state regulation is assessed by examining how cuddly they are in response to being held, how consolable they are while crying, and how well they can quiet themselves while crying; and (g) the social interaction of infants is assessed by examining how well they track both auditory stimuli (e.g., an adult voice) and visual stimuli (e.g., a face).

The NBAS has proven to be a valid and reliable measure of infant behavior in early life and has been used extensively in both research and clinical contexts. Researchers, for example, have used the NBAS

to investigate the effects of maternal substance abuse, cross-cultural differences in infant behavior, the possible developmental outcomes of children, the effects of tactile stimulation on preterm infants, and the effects of obstetric medication on later development. In clinical settings, the NBAS has been administered in the presence of parents to help them better understand the capacities of their infants. By administering the scale in the presence of parents, clinicians can stress the uniqueness of an infant, share concerns about the infant, and foster a positive relationship between the family and themselves. In sum, the NBAS is a valid assessment of infant behavior that has been used widely in both research and clinical contexts.

—Matthew J. Hertenstein and Jennifer L. Porter

Further Reading

Brazelton, T. B., & Nugent, J. K. (1995). *Neonatal Behavioral Assessment Scale* (3rd ed.). London: MacKeith Press.

Neonatal Behavioral Assessment Scale and T. Berry Brazelton description: <http://www.brazelton-institute.com>

NEWMAN-KEULS TEST

Newman and Keuls proposed a procedure for pairwise testing of means in a one-way analysis of variance. The procedure is routinely applied after a significant overall F test. A series of critical differences (CDs) is used to evaluate the significance of the difference between each pair of means. The pairs are investigated systematically from the largest to the smallest difference. CDs are all based on the Studentized range distribution. The largest CD is applied to the difference between the largest and smallest means and is the same as the CD of Tukey's honestly significant difference procedure.

The second largest CD is applied to two differences: (a) the difference between the largest mean and the next-to-smallest mean and (b) the smallest mean and the next-to-largest mean. If there are k means altogether, there will be $k - 1$ CDs. The second largest CD is identical to the CD for Tukey's honestly significant difference, which would be applied to a set of $k - 1$

means. Testing is continued until all pairs are tested, but with the restriction that no pair can be significantly different if that pair is between two means that are not significantly different. This restriction requires careful ordering of the testing of each pair. However, with equal sample sizes, the process is not too difficult.

Even when applied correctly, the original Newman-Keuls procedure has several problems. First, it can have excessively high Type I error rates. Second, it can have lower power (i.e., higher Type II error rates) than procedures that have good control of Type I errors. Third, when it appears to be more powerful than alternative methods, it may actually be less powerful than an alternative that provides equivalent Type I error control. Fourth, it is not suited for data sets with unequal sample sizes, and some approximation must be considered.

In the case of exactly $k = 3$ means, Fisher's least significant difference procedure can be more powerful than the original Newman-Keuls, and least significant difference limits the probability of a Type I error rate to the nominal level α of the test. For $k \geq 4$, Roy Welsch provided a table of modified values from the Studentized range distribution. Those values produce CDs that limit the probability of one or more Type I errors to the level α of the statistical test. The Welsch CDs remove most of the objections to the original Newman-Keuls procedure.

The Welsch modification of the Newman-Keuls can be slightly modified further to give an additional increase in power. The original Newman-Keuls and the Welsch modified Newman-Keuls use the Studentized range statistic to test the full null hypothesis that all k population means are equal. That is, the overall F test is not required for either of these procedures. However, Juliet Shaffer noted that a step-down procedure such as the Newman-Keuls can be modified to test the full null hypothesis with the overall F test of the analysis of variance. In that case, the largest and smallest means are tested for a significant difference with the CD for $k - 1$ means, provided the overall F test is significant. All other pairs are tested as before. This Shaffer-Welsch version of the Newman-Keuls eliminates all Type I error problems in the equal-sample-size case and is more powerful than most other procedures of equal difficulty.

Table 1 A Control and Several Treatment Means

	Control	Treat 1	Treat 2	Treat 3	Treat 4
Mean	4.05	3.25	3.18	3.15	2.24
N	42	24	38	40	40

Note: Treat = treatment group.

All versions of the Newman-Keuls procedure discussed above become excessively complex with unequal sample sizes. However, the harmonic means of sample sizes (illustrated below) covers many empirical data sets. If the largest sample size is no more than twice the smallest sample size, then the harmonic means of sample sizes provides a good approximate solution to the testing of pairwise differences.

Illustrative Example

Consider the following data, in which four treatment groups are being compared to a control. Lower scores indicate better performance.

The within-groups mean square (MS_{WG}) for these data is 3.0326, with degree of freedom (df_{WG}) = 179. The harmonic mean of sample sizes is

$$\tilde{N} = \frac{k}{\frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_k}}$$

For this example, we get

$$\tilde{N} = \frac{5}{\frac{1}{42} + \frac{1}{24} + \frac{1}{38} + \frac{1}{40} + \frac{1}{40}} = \frac{5}{.14179} = 35.2629.$$

Thus, we approximate all five groups as having a sample size of approximately 35.2629. The analysis of variance for these data would give $F = 5.75 > 2.43 = F_{.65}(4,160) > F_{.65}(4,179) = \text{critical value (CV)}$. Therefore, we reject the full null hypothesis at the .05 level and proceed to pairwise testing.

The largest CD is obtained with the Welsch critical value (WCV) and the formula

$$CD = WCV \sqrt{\frac{MS_E}{\tilde{N}}}$$

$$CD = 3.68 \sqrt{3.0326/35.2629} = 3.68 \sqrt{0.8600}$$

$$CD = 3.68(.29325) = 1.07918 = 1.079.$$

For smaller differences, we have

$$CD_2 = 3.64 \sqrt{\frac{MS_E}{\tilde{N}}}$$

$$CD_2 = 3.64(.29325)$$

$$CD_2 = 1.0674 = 1.07$$

$$CD_3 = 3.33 \sqrt{\frac{MS_E}{\tilde{N}}}$$

$$CD_3 = 3.33(.29325)$$

$$CD_3 = 0.97652 = .977.$$

Table 2 shows the differences between ordered means.

Table 2 Difference Between Ordered Means

	Tr 4 2.24	Tr 3 3.15	Tr 2 3.18	Tr 1 3.52	Control 4.05	CD
Tr 4 = 2.24	—	.91	.94	1.28*	1.81*	1.08
Tr 3 = 3.15		—	.03	.37	.90	1.08
Tr 2 = 3.18			—	.34	.87	1.07
Tr 1 = 3.52				—	.53	.975

Note: Tr = treatment group; CD = critical difference.

* $p < .05$

The largest difference is 1.81 between Treatment 4 and the control. This exceeds the largest CD of 1.08 and is therefore significant. The individuals in Treatment 4 have significantly lower average scores than does the control. The two second differences are 1.28 and .90. The value of 1.28 also exceeds 1.08 and is significant. Those in Treatment 4 are significantly lower than are those in Treatment 1. However, .90 is less than the CD of 1.08 and is not significant. There is no significant difference between the mean for Treatment 3 and the control.

At this point, all further testing of the step-down procedure must be done very carefully. The next set of

differences is .94, .37, and .87. The difference of .94 is compared to the CD of 1.07 and is not significant. The mean for Treatment 4 is not significantly different from the mean of Treatment 2. The differences .37 and .87 cannot be significant because they are smaller than .90, which was nonsignificant.

The final four differences, .91, .03, .34, and .53, would ordinarily be compared to the CD of .975 for significance. However, they are nonsignificant without comparison to .975 because they are all contained within a nonsignificant difference.

—Philip H. Ramsey

Further Reading

- Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81, 1000–1004.
- Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods*, 7, 504–523.
- Ramsey, P. H., & Ramsey, P. P. (1990). Critical values for two multiple comparison procedures based on the Studentized range distribution. *Journal of Educational Statistics*, 15, 341–352.
- Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range test. *Journal of Educational Statistics*, 4, 14–23.
- Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, 72, 566–575.

NOMINAL LEVEL OF MEASUREMENT

As its name suggests, nominal (“in name”) level measurement refers simply to labeling or classifying data that belong to different categories within one variable. Some common examples of nominally measured variables in social research include sex, race, political party, religious preference, geographical location, and college major. In these variables, each response fits into one and only one category. A respondent is *either* female *or* male, Republican, Democrat, *or* Independent. To evaluate each of these variables, a researcher

would assign a different number to all categories belonging to the variable. For example, all females might be labeled “1” and all males “2.”

It is important to keep in mind that these values are arbitrary and are only placeholders for longer names. The example above would not mean that males are twice as gendered as females; it simply asserts that female is different from male. The labels easily could be reversed or could be 3s and 9s and still provide the exact same information. Likewise, we would not say that a Native American does not have more race than an African American or that an English student is more majored than a biology student. In fact, the only reason to measure data nominally is if the different categories within one variable *do not* differ in any “direction”—that is to say, we cannot determine which category represents more or less of a trait or variable. The numbers simply help shorten the category names and often are more convenient when researchers use statistical software to analyze data.

Because the only information we can gain from nominal data is “same” or “different,” nominal is considered the “weakest” or least precise level of measurement (followed by ordinal, interval, and then the most precise level, ratio). Beyond the fact that mathematical operations do not apply to them, nominal data are limited because the only appropriate measure of central tendency to use with them is the mode. Similarly, no notion of standard deviation exists with such data. Analysis of nominal data should never include a mean or a *t* test, which would quantitatively compare one category to another.

One should use a more precise level of measurement whenever possible. For instance, when evaluating high school football teams’ performances, finding out how many more points the Mustangs scored per game this season than the Broncos (interval measurement) would provide more useful information than finding out only which team had more points per game but not knowing how many more points (ordinal measurement) or than merely stating that the Mustangs and the Broncos are different teams without specifying a direction that indicates which,

on average, scored more points per game this season (nominal measurement). Only when different responses to an item measuring the same variable *cannot* reveal any directionality is nominal the most appropriate level of measurement to employ.

—Kristin Rasmussen

See also Interval Level of Measurement; Ordinal Level of Measurement; Ratio Level of Measurement

Further Reading

Lane, D. (2003). *Levels of measurement*. Retrieved from <http://cnx.rice.edu/content/m10809/latest/>

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

Sirkin, R. M. (Ed.). (2005). *Statistics for the social sciences*. Thousand Oaks, CA: Sage.

Levels of Measurement page from Research Methods Knowledge Base by W. M. Trochim: <http://www.socialresearchmethods.net/kb/measlev1.htm>

NOMOTHETIC VERSUS IDIOGRAPHIC

The nomothetic (from the Greek term for *lawgiving*) approach to science seeks lawfulness by testing hypotheses. It applies research-supported general formulations to particular cases and uses deductive reasoning. The idiographic (from the Greek term for *oneself, one's own*) approach seeks lawfulness by inspecting individual cases and accounting for them; it uses inductive reasoning. General rules are inferred from individual cases.

The nomothetic approach to assessment interprets measurements and observations by comparing them with group norms. The individual case is understood as an instance of a hypothetical general case. The idiographic approach seeks to understand measurements and observations as a function of the individual's history and environment. The individual case is understood in its own context. For example, a newborn might weigh 5 pounds at birth. Nomothetic data show that this is in the 2nd percentile; since 6% of babies weigh too little for their own good, this weight may be

cause for concern. Idiographic data—which may include behavioral and other physiological observations of the infant, along with information about the due date—can be used to understand whether this particular baby needs intervention. Nomothetic and idiographic data can thus be used integratively.

Psychological science currently favors the nomothetic approach over the idiographic. A study that shows that, in general, siblings placed together in foster care fare better than siblings placed apart is seen these days as more scientific than a study that examines whether a particular sibling pair should be placed together or apart. Psychological assessment is similarly inclined at this time, with behavior analysis and projective testing disfavored in comparison with scores on scales that have meaning only in relation to normative data. The nomothetic approach decreases variation associated with any given psychologist, but it applies to the individual case only roughly. The idiographic approach, conversely, highlights the case at hand but depends for validity on the particular psychologist.

No assessment device or strategy is entirely nomothetic or idiographic. Even personality inventories, whose individual items are ignored by psychologists in favor of scaled scores, show idiographic features when the psychologist considers whether a particular individual is adequately represented by group norms. Sources of inadequate representation might include some distinguishing feature of the individual or of the situation or of the way the test was administered. Can a personality test normed on the general population be used to categorize or describe a woman of Icelandic heritage? A man recently diagnosed with diabetes? Custody litigants?

Even purely idiographic, unnormed assessment techniques, such as asking about childhood memories or about hobbies, reveal nomothetic features when psychologists consider whether their own personal or cultural backgrounds are affecting their interpretations. In this context, cultural competence might be defined as awareness of one's personal norms and their potential inapplicability to other people.

—Michael Karson

Further Reading

- Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Sarasota, FL: Authors Cooperative.
- Ossorio, P. G. (1983). A multicultural psychology. In K. E. Davis & R. M. Bergner (Eds.), *Advances in descriptive psychology* (pp. 13–44). Greenwich, CT: JAI Press.

NONPARAMETRIC STATISTICS

A common statistical problem is to estimate a parameter of the population (such as the mean) or to test whether a parameter is different from some specified value. To carry out inferences, some specific assumptions are made about the distribution of the population; the most common assumption is that the population follows the normal distribution. The various statistical methods developed for these situations are collectively known as *parametric* statistical methods.

Many of the most widely used *nonparametric* procedures provide an alternative to a standard parametric procedure. These nonparametric methods are valid under assumptions more general than those required for parametric inference. In particular, many of these nonparametric procedures are *distribution-free*; they do not require one to make specific assumptions about the form of the distribution of the population. Usually, the only assumption needed to carry out nonparametric procedures is that the population distribution is continuous and, in certain cases, symmetric. In some cases, a nonparametric procedure is not an analogue to a given parametric procedure. This occurs in statistical problems in which one might want to make inferences more general than those involving an individual parameter. For example, there exist nonparametric procedures to test whether two populations have the same underlying distributions or to determine a confidence band for a distribution function.

Advantages of Nonparametric Procedures

When one is presented with the choice between a nonparametric procedure and its parametric counterpart, the former has several advantages. When a

distribution-free nonparametric method exists, exact p values for tests and exact coverage probabilities for confidence intervals can be calculated under fairly general assumptions about the population. Conversely, the p values reported for parametric tests under the assumption of normality are exact only when the population distribution is normal; for other distributions, typically the p values are approximate, with the approximation being better for larger sample sizes. Many nonparametric statistics are relatively simple functions of the ranks of the observations. This implies that one can make inferences about the population without having to know the magnitudes of the sample observations. In addition, when a nonparametric procedure is only a function of the ranks of the observations, the procedure will be insensitive to outliers.

One can compare the performance of a nonparametric method to its parametric counterpart (if one exists) by examining the *asymptotic relative efficiency* of the two techniques. A given estimator of a parameter (say, the sample average for the center of a symmetric population) is more efficient than a competing estimator if the variance of the former is less than the variance of the latter when the sample size is extremely large. When the underlying population distribution is normal, nonparametric analogues of many of the classical procedures based on the assumption of normality are only slightly less efficient. That is, as the sample size gets large, the variance of the nonparametric estimator is not much larger than the variance of the parametric estimator. When the population distribution is not normal, the nonparametric method can be significantly more efficient than the competing parametric method.

Standard Nonparametric Methods

The more common nonparametric procedures used in practice will be described briefly below. Many of these techniques are discussed in more detail in their respective entries in this encyclopedia. The following description of nonparametric procedures is limited to the context of hypothesis testing. For most of the problems discussed here, methods exist to estimate a population parameter and to form a confidence

interval about the parameter. For information on those topics, as well as nonparametric procedures not listed here, the reader is referred to the textbook *Nonparametric Statistical Methods* by Myles Hollander and Douglas Wolfe.

When one is conducting a hypothesis test, nonparametric test statistics are generally calculated by counting, ranking, or both. Counting statistics involve counting the number of observations that exceed a specified value. Statistical tests based on counting statistics are referred to as *sign tests*. For ranking statistics, the observations are ordered from least to greatest. The test statistic is then a function of the ranks of the observations. Tests based on ranking statistics are called *rank tests*. *Signed rank tests* utilize a statistic that involves both counting and ranking.

One-Sample Location Problem

A common problem in statistical inference is to determine whether the center of the population distribution equals some particular value. For normal populations, one uses the mean as a measure of the center and the t test for a population mean for making inferences. In nonparametric statistics, the median (denoted θ) is used as a measure of the center. The question of interest is then whether θ equals some specified value θ_0 . Two nonparametric tests that can be used are the sign test and the Wilcoxon test (also known as the Wilcoxon rank-sum test). For both tests, we assume we have a set of n observations $Z_1, Z_2, \dots, Z_n$ obtained via random sampling from a continuous population. For the Wilcoxon test, we add the assumption that the underlying distribution is also symmetric.

To test the null hypothesis that θ equals θ_0 with the sign test, we count the number of observations greater than or equal to θ_0 . Our test statistic is

$$B = \text{number of } Z_i\text{'s greater than } \theta_0.$$

The total number of observations greater than θ_0 has a binomial distribution with parameters n and p , where $p = P(Z_i \geq \theta_0)$, the probability that Z_i is greater than or equal to θ_0 . This result holds because of the

independence of the observations. Under the null hypothesis that θ_0 is the true population median, $p = 1/2$ (since by definition of the median, one half of the population distribution is above θ_0). Exact p values for the test can then be calculated from this null distribution and will provide a distribution-free test of the null hypothesis that θ equals θ_0 .

If one is willing to impose the assumption that the underlying population distribution is both continuous and symmetric, one can use the Wilcoxon test. This procedure tends to be more efficient than the sign test for testing the null hypothesis that θ equals θ_0 , but is less general since it carries an additional assumption about the population distribution. For our random sample of n observations $Z_1, Z_2, \dots, Z_n$, the test procedure is (a) identify the observations that exceed θ_0 , (b) calculate $|Z_i - \theta_0|$ for each observation, and (c) rank the differences calculated in step (b) from smallest to largest. The Wilcoxon test statistic is then the sum of the ranks of the differences for only the observations that exceed θ_0 .

The theoretical basis of the Wilcoxon test arises from the distribution of ranks. Let R_i denote the rank of the i th value of $|Z_i - \theta_0|$. The set of ranks $R_1, R_2, \dots, R_n$ are a specific permutation of the integers from 1 to n ; there are $n!$ possible permutations. The set of ranks has a jointly uniform distribution, where the probability of a particular set of ranks is $1/n!$. Moreover, whether a particular observation is greater than or equal to θ_0 is independent of the magnitude of the difference between that observation and θ_0 when the observations are taken from a symmetric distribution. Thus, the Wilcoxon test is distribution-free with a null distribution that can be found by means of the uniformity of the ranks, the binomial nature of the counting, and the independence of the counting and ranking procedures. Tables of the null distribution are available in most nonparametric texts.

Both the sign test and the Wilcoxon test can be applied to problems involving *paired replicates data*; paired replicates data would occur if two measurements were taken from the same group of participants at two different times. For example, one can measure opinions about a political candidate before and after a

debate or patients' medical condition before and after a treatment. One generally wishes to test whether the characteristic in question changed systematically from the first measurement to the second. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ represent our n pairs of measurements, where X_i represents the first measurement and Y_i represents the second measurement. Then we can define Z_i to equal $(Y_i - X_i)$ and carry out the sign or Wilcoxon test using these Z_i s. In parametric inference, the standard approach to this type of problem is to use the paired samples t test.

Two-Sample Problems

Suppose we have two independent random samples from two different populations. Several nonparametric procedures exist to test whether the two underlying population distributions are identical against the alternative that one population tends to have larger values than the other.

The Mann-Whitney U test (Wilcoxon rank-sum test) is the most commonly used nonparametric procedure for this problem. Its comparable test in parametric inference is the t test for two population means. Let $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ denote our two independent random samples. Note that the sample sizes need not be equal. We assume that the samples were taken from populations with continuous distributions and that the distributions of both populations have the same shape. The only difference between the two population distributions is that one may be a shifted version of the other. The shift in the distribution may be attributable, for example, to one group's receiving a medical treatment and another group's being a control group. To test the null hypothesis that there is no shift in the distributions, we first rank the combined group of $m + n$ observations. The test statistic is then

$$W = \text{sum of the ranks for the } Y_i \text{ observations.}$$

Since this test statistic involves the ranking of observations, the theoretical basis for the null distribution is similar to that used for the Wilcoxon test. Here, there are $\binom{m+n}{n}$ possible sets of ranks for Y ,

each equally likely, and the Wilcoxon test is distribution-free.

Let U denote the test statistic for the Mann-Whitney U test. Then U can be computed from W by the relationship

$$W = U + \frac{n(n+1)}{2}.$$

The two tests are therefore equivalent. Tables of the null distribution of W or U are available in most nonparametric texts.

Other nonparametric tests exist for detecting differences in population distributions. One popular test is the Kolmogorov-Smirnov test for two samples. This procedure tests the null hypothesis that the population distributions are identical against the null hypothesis that the distributions differ at one or more points. Tests also exist for detecting differences in the dispersion parameter of two distributions, including one based on the *jackknife*.

Other Statistical Problems

Nonparametric procedures exist to address numerous other statistical problems. For one-way analysis of variance problems, several distribution-free methods exist, and the choice depends on the alternative hypothesis of interest. To test the null hypothesis that all treatment effects are equal against the alternative that at least two treatment effects differ, the most common nonparametric procedure is the Kruskal-Wallis one-way analysis of variance. The Jonckheere Terpstra test is applicable when the alternative hypothesis is that the treatments are ordered. For the two-way layout, the Friedman two-way analysis of variance is a nonparametric alternative to the two-way layout procedure based on the normal distribution.

Two nonparametric methods are commonly used to test for the independence between two variables from a bivariate distribution. Kendall's tau statistic allows for a nonparametric test based on the number of concordant and discordant pairs of observations. Spearman's rank correlation is an alternative to the Pearson product-moment correlation, which uses the

ranks of the observations in place of the actual data values.

The *bootstrap* is a computationally intensive technique that can be used to obtain confidence intervals for complicated estimators. The procedure resamples from the data to approximate the standard error of the estimator. The bootstrap is nonparametric in that it does not require any assumptions about the form of the underlying distribution.

Nonparametric methods have also been adapted to regression problems. One approach is to use traditional least squares procedures for estimation but remove the assumption of normality of the errors when making inferences. A significant amount of recent nonparametric research has been devoted to *nonparametric regression*. This approach does not assume a prior regression model; rather, the data are used to develop the regression relationship. Nonparametric regression relies on computational methods to “smooth” the data and does not generally utilize ranking procedures.

Computer Software

Most of the major software packages for statistical analysis can perform nonparametric procedures. Minitab Release 14 can conduct one-sample sign and Wilcoxon tests, the Wilcoxon rank-sum test, and analysis of variance using the methods of Kruskal-Wallis and Friedman. These procedures are generally based on large sample approximations. SAS Version 9, StatXact, and SPSS Exact are three packages that can compute exact p values and coverage probabilities. For SAS, the default is to use approximations, but the user can specify an exact test in the option statement. StatXact is a specialized software package available from the Cytel Corporation for nonparametric and categorical data analysis. It is available as a stand-alone package or as a series of PROCs that can be used in SAS. The Web site for Cytel at the end of this entry compares the procedures that can be performed by StatXact, SPSS Exact, and SAS.

—Christopher J. Sroka

See also Binomial Distribution/Binomial and Sign Tests; Bivariate Distributions; Confidence Interval; Hypothesis and Hypothesis Testing; Kendall Rank Correlation; Kolmogorov-Smirnov Test for Two Samples; Kruskal-Wallis One-Way Analysis of Variance; Mann-Whitney U Test (Wilcoxon Rank-Sum Test); Median; One-Way Analysis of Variance; Paired Samples t Test (Dependent Samples t Test); Pearson Product-Moment Correlation Coefficient; Random Sampling

Further Reading

- Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transaction of the Royal Society of London*, 27, 186–190.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing* (2nd ed.). New York: Dekker.
- Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference* (4th ed.). New York: Dekker.
- Hajek, J. (1969). *A course in nonparametric statistics*. San Francisco: Holden-Day.
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
- Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 7, 29–43.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.
- Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.
- Savage, I. R. (1953). Bibliography of nonparametric statistics and related topics. *Journal of the American Statistical Association*, 48, 844–898.

StatXact compared to SPSS Exact Tests and SAS software: http://www.cytel.com/Papers/SX_Compare.pdf

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Du, Y., Akritas, M. G., Arnold, S. F., & Osgood, D. W. (2002). Nonparametric analysis of adolescent deviant behavior. *Sociological Methods & Research*, 30(3), 309–340.

There are many different models of quantitative analysis, parametrics and **nonparametrics** just

being two. In this study, Yunling Du and his colleagues studied the impact of certain personal characteristics on marijuana use while controlling for routine everyday activities. The assumptions underlying several logistic regression models, as well as the parametric analysis of covariance (ANCOVA) model, are violated in this example. More critical from the practical point of view is the fact that the covariate is measured on an ordinal (noninterval) scale, implying that the results from such analyses depend on the chosen scale. The analysis of main effects and interactions is compared with similar analyses by means of the parametric ANCOVA and logistic regression models. The formal analysis is supplemented by new exploratory data analysis plots. Significance tests for certain ordered and other patterned alternatives are proposed and evaluated via simulations. Results are compared with those from logistic regression models treating the levels of the factors as ordinal.

NONPROBABILITY SAMPLING

Nonprobability sampling includes several versions of survey sampling that often are expedient to implement but do not allow calculation of the probability of selection of the sample from among possible samples from a population. The goal of a survey is to gather data in order to describe the characteristics of a population. A population consists of units, or elements, such as individuals, households, businesses, tracts of land, or inventory records. A survey collects information on a sample, or subset, of the population. In some surveys, specific units or elements are chosen by the survey designers to be in the sample. Interviewers are assigned to interview the members of or gather data on the units in the selected sample. Sometimes multiple attempts at contacting and collecting data from the selected sample members are made. In many situations, it is possible to compute the probability that any member of the population is in a sample of a certain size selected with a specified design or protocol. It also is possible in many situations to compute the

probability overall that the sample is selected from the population. Such a design is a *probability sampling* design. Examples of probability sampling include simple random sampling, stratified random sampling, and cluster sampling.

In *nonprobability sampling* designs, it is not possible to compute the probabilities of selection for the samples overall or, usually, for individuals. Examples of nonprobability sampling include convenience sampling and quota sampling. In convenience sampling, interviewers themselves are given some latitude in selecting the population members to interview or the units on which to record data. That is, the survey designers and planners do not strictly control the selection of the sample. Convenience sampling occurs in many forms, including selection by an interviewer of people at a shopping mall, selection by a waiter of customers at a restaurant, and Internet and call-in polls. Quota sampling is similar to convenience sampling but requires interviewers to collect certain data on certain numbers of individuals from each of several population subgroups or strata. For example, a quota sample could require interviewers to interview a specified number of females and males within age groups in administrative regions within a state. One could implement such a study by randomly calling phone numbers from a telephone book, asking to speak to a person in the household who belongs to one of the available strata, and conducting an interview if possible. If no one is available at a household in an allowable stratum, then the interviewer simply calls the next telephone number.

Estimates of population characteristics based on nonprobability samples are affected by *selection bias*. Since the interviewers choose respondents that they want to interview, there is a potential for selection bias. If the respondents in the survey are systematically different from the general population on the variables being measured, then estimates of characteristics will be different on average from what they would have been with a controlled probability-sampling scheme. In probability sampling, the survey planner or researcher controls which units are in the sample and selects the sample using known

probabilities of selection. The probabilities of selection can be used to produce estimates of population characteristics without the problem of selection bias.

Probability sampling is the standard methodology for large-scale surveys intended to support scientific studies and decision making for government policy. Nonprobability sampling, on the other hand, is quite common in marketing surveys and less formal studies. Nonprobability sampling certainly can produce useful information for some purposes. One attempt to adjust for the fact that probabilities of selection are unknown is to use weights, usually called *survey weights*, in analysis. These weights are computed so that the weights for sampled individuals in a particular stratum sum to a number proportional to the actual number of people known to exist in the population in the stratum. Although this adjustment can help make the sample more representative of the population in analysis, it cannot overcome the fact that there could be a remaining bias due to noncontrolled random selection of the sample.

—Michael D. Larsen

See also Convenience Sampling; Quota Sampling

Further Reading

Guo, S., & Hussey, D. L. (2004). Nonprobability sampling in social work research: Dilemmas, consequences, and strategies. *Journal of Social Service Research*, 30, 1–18.

NORMAL CURVE

A normal curve is the graph of the probability density function (PDF) for a normal distribution (see Figure 1). It is not just one but a family of curves of the same general form characterized by two parameters: the location parameter, or mean, and the scale parameter, or standard deviation. The formula for this family of curves is



Figure 1 Deutsche Mark Bill With Gauss and Bell-Shaped Normal Curve

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < +\infty,$$

where

mean $\mu \in (-\infty, +\infty)$,

standard deviation $\sigma \in (0, +\infty)$.

For a continuous random variable X that is normally distributed with parameters μ and σ^2 , the distribution is usually denoted by $X \sim N(\mu, \sigma^2)$.

A Brief History of the Normal Curve

The discovery of the normal curve, also known as the “bell-shape” curve or the Gaussian curve, can be dated to the 17th century, when Galileo Galilei, an Italian physicist and astronomer, noted that the measurement errors in astronomical observations were very systematic and that small errors were more likely to occur than large errors. In 1778, Pierre-Simon Laplace, while working on his famous *central limit theorem*, noted that the sampling distribution of the sample mean approximated a normal distribution and that the larger the sample size, the closer the distribution would be to a normal distribution, no matter what the population distribution might be. Also in the 18th century, a French statistician, Abraham de Moivre, who was often asked to do statistical consulting for gamblers, found that when the number

of events (e.g., coin flips) increased, the shape of the binomial distribution would approximate a symmetrical and smooth curve. However, the mathematical formula for this curve was not discovered until the 19th century, by Adrian Marie Legendre in 1808 and Carl Friedrich Gauss in 1809. The German 10 deutsche mark bill (see Figure 1) had Gauss's picture on it, along with the well-known bell-shaped normal curve and its formula.

Important Properties of a Normal Curve

This family of curves has the following characteristics, which are important to know. For those who are unfamiliar with the calculus calculation, all functions involving integration can be safely skipped. Graphical illustrations are used to facilitate an understanding of the concepts presented, and knowledge of basic algebra is assumed.

1. The mode, the mean, and the median are all at the same point on the abscissa, the horizontal axis of the curve (see Figure 2). That is to say, mode = mean = median for a normal distribution.

2. The curve is symmetrical about the point on the abscissas that denotes the mean, the mode, or the median, with equal numbers of observations above and below the point (see Figure 2).

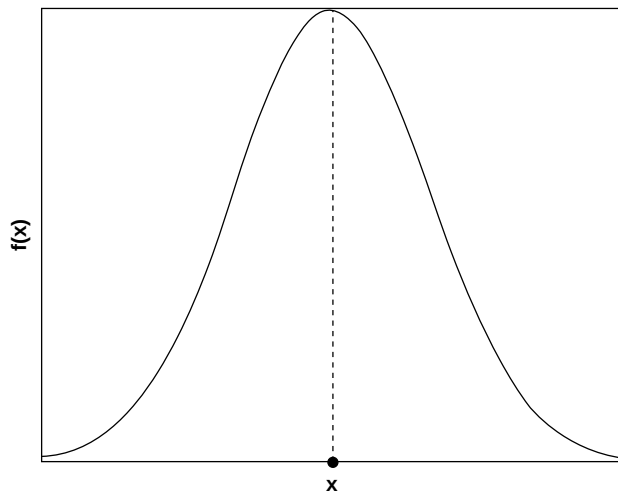


Figure 2 Typical Normal PDF Curve

3. The skewness and the kurtosis for a normal distribution are both 0. However, statisticians who omit the subtraction of 3 from the kurtosis formula will report the kurtosis value to be 3 for a normal distribution.

4. The area jointly determined by two points (a, b) on the abscissa and the normal probability density curve indicates the probability that an observation falls within the intervals of $[a, b]$, $[a, b)$, $(a, b]$, or (a, b) . In other words,

$$\begin{aligned} p(a \leq X \leq b) &= p(a \leq X < b) = p(a < X \leq b) \\ &= p(a < X < b) = \int_a^b f(x)dx. \end{aligned}$$

Note that the probability associated with a particular value of a continuous variable is 0. For example, when $X \sim N(\mu, \sigma^2)$, $p(X = a) = 0$. Thus,

$$\begin{aligned} p(a \leq X < b) &= p(X = a) + p(a < X < b) \\ &= 0 + p(a < X < b) \\ &= p(a < X < b). \end{aligned}$$

5. When $a \rightarrow -\infty$, $b \rightarrow +\infty$, the probability mentioned in Paragraph 4 approaches 1. In other words, the total probability under the normal curve is always 1, or

$$\int_{-\infty}^{+\infty} f(x)dx = 1,$$

which also indicates that the probability of a sure event that $X \in (-\infty, +\infty)$ is 1.

6. Based on the probability theory, the cumulative distribution function (CDF) $F(x)$ is defined as the probability that a random variable X takes on a value that is less than or equal to a given x . For $X \sim N(\mu, \sigma^2)$, the corresponding CDF is calculated as follows:

$$\begin{aligned}
 F(x) = p(X \leq x) &= \int_{-\infty}^x f(t) dt \\
 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.
 \end{aligned}$$

Note that the normal PDF is substituted for $f(t)$ in the formula. A normal PDF, along with its CDF, is illustrated in Figure 3, with the solid curve representing the PDF and the dashed curve the CDF. The area under the normal probability density curve when $x \leq a$ is equal to the value $F(x)$ assumes when $x = a$, or $F(a)$ in the normal CDF curve. This is an important property of the normal curve in that it relates the normal PDF to the normal CDF.

7. The normal curve indicates that all values of x are normally distributed with mean μ and standard deviation σ . In a statistical analysis, it is not uncommon that some normal random numbers have to be generated. In that case, those normal random numbers should be generated from the x -axis in a normal curve by means of a random number table or a computer software package.

8. When $X \sim N(0,1)$ or $\mu = 0$, $\sigma = 1$, the normal distribution is known as the standard normal

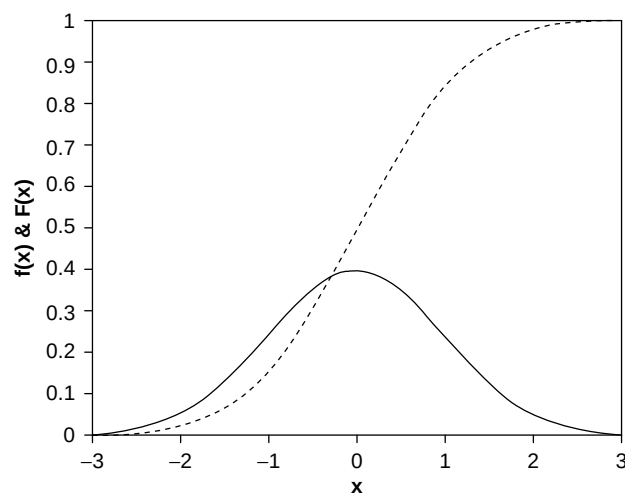


Figure 3 Normal PDF and CDF Curves

distribution. The standard normal curve is centered at 0, and the standard deviation is 1.

9. Any normal distribution $X \sim N(\mu, \sigma^2)$ can be transformed to a standard normal distribution by the following formula:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

When standardization is performed, the normal curve will change its shape, with standard deviation increasing or decreasing to 1, and its location on the abscissa, with the midpoint of the normal curve moving to 0.

10. If the normal curve is divided into standard deviation units, a known proportion of observations fall within each portion of the curve. An empirical rule states that

- About 68.26% of the observations fall within 1 standard deviation below and above the mean.
- About 95.44% of the observations fall within 2 standard deviations below and above the mean.
- About 99.73% of the observations fall within 3 standard deviations below and above the mean.

To find out the proportion of observations that fall under a normal curve, not necessarily the standard normal one, between any arbitrary pair of points, we can consult a table containing the areas under a standard normal curve. Such a table can be found at the end of most statistics books as an appendix. Software packages containing modules or procedures for computing the areas are also widely available.

A Typical Problem

A typical example is illustrated below to demonstrate what people usually do with a normal curve. Notice that, for a normal curve, what is usually of interest is not the value of the function $f(x)$ but the area, or the probability, under the curve.

Suppose that the grades in a statistics exam have an $X \sim N(70, 10^2)$ distribution and that the grading scale is as follows:

91 to 100: Excellent

60 to 90: Good

Below 60: Fail

What is the proportion of students passing the exam?

Solution:

Figure 4 is a graphical illustration of this problem.

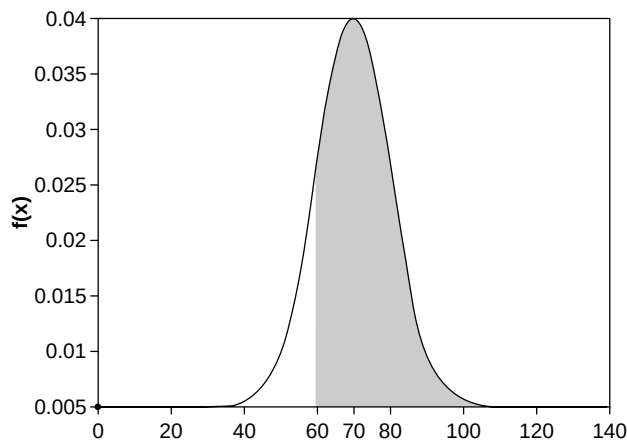


Figure 4 Normal PDF Curve for the Grades

The proportion of students passing the exam is the area under the curve between $x = 60$ and $x = 100$, which is indicated by the shaded area in Figure 4.

$$\begin{aligned}
 p(\text{Pass}) &= p(\text{Good}) + p(\text{Excellent}) \\
 &= p(60 \leq X \leq 100) \\
 &= p\left(\frac{60 - 70}{10} \leq Z = \frac{X - 70}{10} \leq \frac{100 - 70}{10}\right) \\
 &= p(-1 \leq Z \leq 3) = 0.84.
 \end{aligned}$$

An alternative method is to use the probability of *Fail* to solve this problem:

$$\begin{aligned}
 p(\text{Pass}) &= 1 - p(\text{Fail}) = 1 - p(X < 60) \\
 &= 1 - p\left(Z = \frac{X - 70}{10} < \frac{60 - 70}{10}\right) \\
 &= 1 - p(Z < -1) = 0.84.
 \end{aligned}$$

It should be noted that, in the solution to the problem, the area under a nonstandard normal curve, as indicated by the shaded area in Figure 4, is expressed in terms of the area under a standard normal curve by the transformation

$$Z = \frac{X - \mu}{\sigma}.$$

Although the two areas are different in shape, they are equal in size, indicating that the probabilities are equal. After the transformation is performed, the standard normal areas table could be employed to find the corresponding probability under the standard normal curve, which should be equal to the probability under the original normal curve.

Impact on Psychological Measurement

The importance of the normal curve can never be emphasized enough in the field of measurement and statistics. First, many variables in the science of psychological measurement are at least approximately normally distributed, and the approximation is very close. Measures of reading ability, job satisfaction, and memory are only a few examples in this field. Second, most statistical tests work well even if the distributions are only approximately normal. Third, assuming normality simplifies the mathematical procedures required to compute probabilities.

Caveats

1. Although the normal curve appears to be bell shaped, not all bell-shaped curves come from a normal distribution. A case in point is the Student t distribution, which is the ratio of a standard

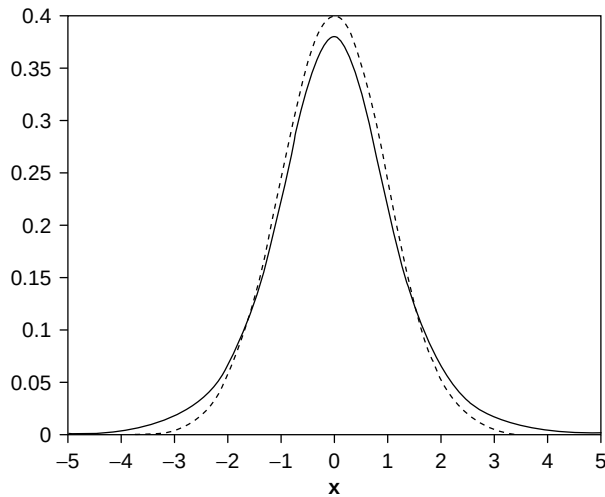


Figure 5 Comparison of a Normal and a t PDF Curve

normal random variable to the square root of an independently distributed chi-square random variable divided by its degrees of freedom. The t PDF also features a family of curves depending on only one parameter: the degrees of freedom ν , with each one of them being bell shaped and symmetric around 0. Figure 5 compares a Student t distribution with $\nu = 5$ (solid line) to the standard normal distribution (dashed line). As can be seen, the curve for the t distribution appears to be very similar to that for a normal distribution.

2. Although most normal curves appear to be moderately sharp and moderately spread out, like the one in Figure 2, some normal curves could be the other way around, depending on the value of the scale parameter or standard deviation σ . Figure 6 shows two of them that do not fit the stereotype of a normal curve. As can be seen, as the value of σ goes down, the steepness of the normal curve goes up. Observations cluster more and more in the middle of the distribution. The steeper normal curve, on the top, has a relatively small σ value of 1 while the normal curve on the bottom is the opposite case, with its σ value being 8 times as large, hence a flat curve. However, although one normal curve may be different from another in terms of steepness, the total area under the curve, or the probability that x assumes a value between negative infinity and positive infinity, or the probability of a sure event, or

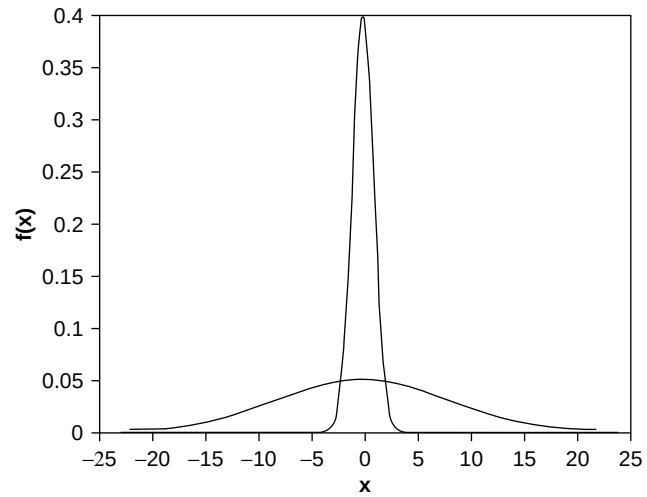


Figure 6 Comparison of a *Steep* and a *Flat* Normal PDF Curve

$$\int_{-\infty}^{+\infty} f(x) dx,$$

is always 1.

3. Nowadays, most computations involving a normal curve or a normal distribution can be performed with a computer software package. Generally speaking, five types of problems are related to this topic.

- Find the cumulative probability p that a value from a normal distribution with specified mean and standard deviation is equal to or less than a specified value x_0 . This type of problem is often related to the concepts of percentile and quantile.
- Find the value from a normal distribution with specified mean and standard deviation at which the cumulative probability is a specified value p_0 . This type of problem also is often related to the concepts of percentile and quantile.
- Find the value of the PDF of a normal distribution with specified mean and standard deviation at a specified value x_0 .
- Generate random numbers from a normal distribution with specified mean and standard deviation.
- Plot a normal probability density curve with specified mean and standard deviation.

Many available software packages could be utilized for these five types of problems. Table 1 summarizes the packages.

Table 1 Related Functions Available in Four Popular Software Packages

Names	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5
Excel	NORMDIST	NORMINV	NORMDIST	N/A	CHART
SPSS	CDF.NORMAL	IDF.NORMAL	PDF.NORMAL	RV.NORMAL	N/A
SAS	CDF	QUANTILE	PDF	RAND	PROC PLOT
Matlab	normcdf	norminv	normpdf	normrnd	plot

For some of the most popular software packages, Table 1 provides the names of functions, modules, and procedures that can be applied to the five very typical problems related to the normal curve and the normal distribution. Note that none of the names is case sensitive, with the exception of Matlab functions and modules, which have to be used in lowercase letters.

4. The family of normal curves that are bell shaped is only for the univariate case, in which only one variable x is involved. However, in the case of multivariate data analysis, the multivariate normal model, which extends the univariate normal distribution model, is commonly used. One example is a bivariate normal distribution model, which applies to two variables. In that case, the bell-shaped normal curve becomes a bell-shaped surface in three dimensions. Accordingly, the probability is indicated by the volume under the bivariate normal distribution surface.

—Hongwei Yang

Further Reading

- Aron, A., Aron, E. N., & Coups, E. (2005). *Statistics for psychology* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Johnson, D. E. (1998). *Applied multivariate methods for data analysts*. Pacific Grove, CA: Duxbury.
- Martinez, W. L., & Martinez, A. R. (2001). *Computational statistics handbook with Matlab*. New York: Chapman & Hall.
- Patel, J. K. (1982). *Handbook of the normal distribution*. New York: Dekker.
- Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall.

Normal curve information:

http://en.wikipedia.org/wiki/Normal_distribution

<http://mathworld.wolfram.com/NormalDistribution.html>
<http://www.answers.com/topic/normal-distribution>
<http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html>
http://www.tushar-mehta.com/excel/charts/normal_distribution
<http://www-stat.stanford.edu/~naras/jsm/FindProbability.html>

NULL HYPOTHESIS SIGNIFICANCE TESTING

Null hypothesis significance testing (NHST) dominates experimental and correlational methods in psychological research. Investigators are typically concerned with demonstrating the existence of an effect, that is, systematic variation in the data that can be distinguished from random noise, sampling error, or variation due to uncontrolled or nuisance variables. The null hypothesis is often, but does not have to be, identified with chance, and a p value is computed to express how improbable observed empirical data are under the assumption that the null hypothesis is true. When this probability falls below the conventional value of .05, it is concluded that the null hypothesis is false and that it is safe to presume the presence of a systematic source of variation. This inference is not strictly logical because *modus tollens* is not valid when stated probabilistically: From the statement “If the null hypothesis is true, then extreme data are improbable,” it does not follow that “If the data are probable, the null hypothesis is false.” Because NHST is a method of inductive, not logical, inference, researchers nevertheless believe that the rejection of the null hypothesis indicates the presence of an effect. In the long run, the argument goes, decisions reached

by NHST will generate knowledge faster than would guessing or doing nothing.

Variants of NHST have been developed by various, and sometimes warring, schools of statistical thought. These schools differ in the assumptions they make about the nature of the data and the hypotheses and about how to make inferences. The following illustrations of possible inference strategies begin with information- and assumption-rich scenarios and proceed to the more degraded scenarios typical of most psychological research.

Full-Suite Analysis

Suppose extensive testing has revealed that average self-esteem scores are $\mu = 68$ and 72 for women and men, respectively, and that the standard deviation within each gender is $\sigma = 20$. A sample of 200 scores with a mean of 71 is drawn from one of the two populations. The null hypothesis H_0 is that women were sampled, and the alternative hypothesis H_1 is that men were sampled. Analysis begins with the calculation of the probability of obtaining a mean of 71 or higher if H_0 is true. The z score for the sample mean is

$$(71 - 68)\sqrt{200}/20 = 2.12,$$

and the probability of a score at least this extreme is .017.

Evaluation of the data under the alternative hypothesis H_1 yields $z = .71$, $p = .24$. That is, the data are not improbable under the assumption that men were sampled. The likelihood ratio (LR) of the two p values, $p(D|H_1)/p(D|H_0)$, is 14.12 , meaning that it is more than 14 times more likely that a sample of men rather than women would yield data of the kind found in the empirical sample. But how likely is it that the sample consisted of men? It is necessary to be explicit about the prior probability of sampling men. A simple intuition is that women and men were equally likely to be sampled, that is, $p(H_0) = p(H_1) = .5$. The summed

products of these prior probabilities and their respective p values is the overall probability of the observed data. Here, $p(D) = p(H_0)p(D|H_0) + p(H_1)p(D|H_1) = .13$. This probability is critical for the calculation of the probability of the null hypothesis given the observed data. Bayes' theorem gives $p(H_0|D)$ as $p(H_0)p(D|H_0)/p(D) = .07$. Because the prior probabilities of the two hypotheses are the same, the ratio of the two posterior probabilities is the same as the LR. It can now be said that the sample is more than 14 times more likely to comprise men than women. The assumption of equal priors was just that, an assumption. Suppose the researcher knew that self-esteem scores were collected at four different sites, only one of which comprised men. Now $p(H_0|D) = .18$, meaning that it is only 4.7 times more likely for the sample data to come from men than from women. Although the prior probability that men were sampled was low, the evidence is still strong enough to reject the null hypothesis that women were sampled and to accept the alternative.

Now consider a study in which 200 women and 200 men are sampled. The null hypothesis is that there is no gender difference in average self-esteem scores ($H_0: \mu_{\text{women}} = \mu_{\text{men}} = 70$, $\sigma = 20$), and the alternative is that there is a 4-point difference ($H_1: \mu_{\text{women}} = 72$, $\mu_{\text{men}} = 68$, $\sigma = 20$). During the early phase of the research program, the two hypotheses may appear to be equally likely to be true. If the gender difference in the sample means is 3.5 points, the revised probability of the null hypothesis is $p(H_0|D) = .09$. As evidence accumulates, researchers become aware that some hypotheses are riskier than others. Suppose gender differences in self-esteem have become well established, so that $p(H_0) = .1$. Now a 3.5 gender difference still renders the null hypothesis less probable ($p(H_0|D) = .01$), but there is less room to move.

These examples are idealized: The properties of the two populations (μ and σ) are known, and credible estimates of their prior probabilities are available. Scientific research must often proceed without this full suite of information. Researchers handle the lack of information by suspending certain kinds of inference or by making defensible assumptions where

good information is missing. If the prior probabilities of competing hypotheses are unavailable in a quantifiable and agreed-on format, they can sometimes be estimated on the basis of prior research or derived from theory.

Power

Many researchers are careful to situate their findings within the context of relevant empirical or theoretical work but refrain from making explicit estimates for their hypotheses to be true. Suppose again that a 3.5 gender difference in self-esteem is found. Evaluation of the data under the two hypotheses yields $p(D|H_0) = .04$ and $p(D|H_1) = .401$, and thus $LR = 10$. With prior probabilities barred from quantitative inferences, researchers can still estimate their study's statistical power to detect a 4-point gender difference. The power of the study is the probability that the null hypothesis will be rejected if it is indeed false. To obtain this probability, it is necessary to find the minimum gender difference leading to the rejection of H_0 . This difference is given by the product of the z score at which $p(D|H_0) = .05$ and the standard error of the difference. (The standard error of the difference between means $\bar{X}_1$ and $\bar{X}_2$ is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2};$$

here, $1.65 \times 2 = 3.3$.)

The power of the study is the complement of the probability of such an effect under the alternative hypothesis. Here, $1 - p(D|H_1) = .64$. In other words, the prior odds that this study would detect an existing difference of 4 points were about 5 to 3.

In principle, many researchers agree that the null hypothesis should not be rejected when $p(D|H_0) > .05$. In practice, however, they tolerate a good number of exceptions, thus opening the door to the murky world of "marginal significance." Researchers typically care more about limiting the probability that a true null hypothesis is rejected than about increasing the

probability that a true effect is detected. Designing a study with a power of .8 is a widely held but seldom attained ideal. One reason for this shortfall is that power consumes resources. In the present example, a total sample of 1,152 individuals would be required to reach the ideal.

Making Decisions

In the decision-theoretic school of hypothesis testing, $p(D|H_0) = .05$ signifies the probability with which a true null hypothesis is rejected. This decision outcome constitutes a Miss (M). Conversely, $p(D|H_1)$ is the probability that a false null hypothesis is not rejected, a circumstance called a False Positive (FP). The complement of an M is a Hit (H), that is, the retention of a true null hypothesis; the complement of an FP is a Correct Rejection (CR), that is, the rejection of a false null hypothesis. The probability of CR is the power of the study. If there are no resources to increase power, it is tempting to admit more FP. If the null hypothesis is rejected with $p(D|H_0)$ as high as .10, power increases from .64 to .76. The practice of adjusting $p(D|H_0)$ is frowned on, however, when it reflects, not the state of the field and thus appropriate prior probabilities, but rather the researcher's desire to obtain significant results.

Without prior beliefs, there is no way of estimating how probable the four outcomes are. It is only possible to state the conditional probability of $p(H)$ relative to $p(M)$ and of $p(FP)$ relative to $p(CR)$. To illustrate what can be gained from estimating the prior of the null, consider $p(H_0) = .75$, .5, and .25. The top panel of Figure 1 shows the four conditional probabilities obtained in a high-powered study. Each quadrant of the bottom panel gives three unconditional probabilities that are obtained as products of the conditional probabilities and the prior probabilities of the hypotheses. When, as in the typical empirical case, the probability of rejecting a true null hypothesis (M) is smaller than the probability of accepting a false one (FP), any decrease in the prior probability of H_0 decreases the overall probability of correct decisions

		Decision Regarding H_0	
		True	False
Reality of H_0	True	Hit $1-p(D H_0) = .95$	Miss $p(D H_0) = .05$
	False	False Positive $p(D H_0) = .20$	Correct Rejection $1-p(D H_1) = .80$

		Reality of H_0		
		True	False	
Reality of H_0	True	$1-p(D H_0)P(H_0)$.7125 .475 .24	$p(D H_0)P(H_0)$.0375 .025 .0125	$P(H_0)$.75 .5 .25
	False	$p(D H_1)P(H_1)$.05 .1 .15	$1-p(D H_1)P(H_1)$.2 .4 .6	$P(H_1)$.25 .5 .75

Figure 1 A Decision-Theoretic Scheme for Null Hypothesis Significance Testing

(here, $p(H) + p(CR) = .91$, $.88$, and $.84$, respectively, for $p(H_0) = .75$, $.5$, and $.25$). This is an odd, but logical, result. As an area of research becomes more mature, null hypotheses become less probable, and the typical conservatism of decision making (i.e., power $< 1 - \text{desired significance level}$) makes it more likely that true effects are missed. Failures to replicate then accumulate, not because previously demonstrated phenomena do not exist, but because studies lack power. Hence, even the principled use of NHST delays scientific progress. The data of solidly designed but underpowered studies are dismissed for ad hoc reasons, or worse, they are seen to add up to a store of anomalies that potentially undermines hard-won knowledge.

Filling In

When a new area of research opens up, it is marked by great uncertainty. The null hypothesis may not have a defensible prior probability, and there may not be a well-formulated alternative hypothesis. Without being able to estimate the posterior probability of the null and with no opportunity to estimate the power of the study, researchers seek to collect only enough data to reject the null. When they do, they can declare only that an effect has been found, and they can report its size (for example with Cohen's d or Pearson's r). Power analyses can be performed with the obtained effect size, but the wisdom of this practice is a matter of debate. Nevertheless, when enough empirical effect sizes have been reported to justify their aggregation by meta-analysis, these combined effect sizes can serve as point-specific research hypotheses for replication and extension studies.

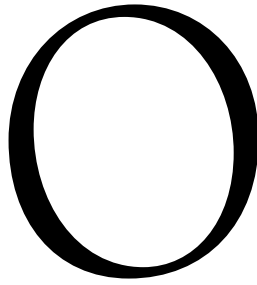
The life course of a typical research area entails a paradox. In the early stages, NHST can be performed only in its most rudimentary form. At this stage, misinformation and miseducation are most likely to contribute to fallacious conclusions, such as the widespread belief that the p value of the data signifies the improbability of the null hypothesis. In the late stages, when specific alternative hypotheses are available, when the power of a study can be determined, and when the probability of hypotheses can be estimated, new data contribute little incremental knowledge. Although NHST can then be used with great precision, its purpose is now to produce judgments about the acceptability of the data and not about the truth or falsity of the hypotheses.

—Joachim I. Krueger

Further Reading

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606.

- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19–24.
- Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, 56, 16–26.
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27, 313–376.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.



I have had my results for a long time: but I do not yet know how I am to arrive at them.

—Carl Friedrich Gauss

O'BRIEN TEST FOR HOMOGENEITY OF VARIANCE

The homogeneity of variance assumption is one of the critical assumptions underlying most parametric statistical procedures, such as the analysis of variance (ANOVA), and it is important to be able to test this assumption. In addition, showing that several samples do not come from populations with the same variance is sometimes of importance per se. Among the many procedures used to test this assumption, one of the most sensitive is the O'Brien test, developed by Ralph G. O'Brien. The null hypothesis for this test is that the samples under consideration come from populations with the same variance; the alternative hypothesis is that the populations have different variances.

Compared with other tests of homogeneity of variance, the advantage of the O'Brien test resides in its versatility and its compatibility with standard ANOVA designs. It is also optimal because it minimizes both Type I and Type II errors. The essential idea behind the O'Brien test is to replace, for each sample, the original scores by transformed scores such that the transformed scores reflect the variance of the sample. Then, a standard ANOVA based on the transformed scores will test the homogeneity of variance assumption.

Motivation and Method

Several tests are available for detecting whether several samples come from populations having the same variances. In the case of two samples, the ratio of the population estimates (computed from the samples) is distributed as a Fisher distribution under the usual assumptions. Unfortunately, there is no straightforward extension of this approach to designs involving more than two samples. By contrast, the O'Brien test is designed to test the homogeneity of variance assumption for several samples at once and with the versatility for ANOVA designs, including contrast analysis and analysis of subdesigns.

The main idea behind the O'Brien test is to transform the original scores so that the transformed scores reflect the variation of the original scores. An ANOVA on the transformed scores will then reveal differences in the variability (i.e., variance) of the original scores, and therefore this analysis will test the homogeneity of variance assumption. A straightforward application of this idea will be to replace the original scores with the absolute value of their deviation to the mean of their experimental group. So, if we denote by Y_{as} the score of participant s in experimental condition a whose mean is denoted by M_a , this first idea amounts to transforming Y_{as} into v_{as} as follows:

$$v_{as} = |Y_{as} - M_a|.$$

This transformation has the advantage of being simple and easy to understand, but unfortunately, it creates some statistical problems (i.e., the F distribution does not model the probability distribution under the null hypothesis), and in particular, it leads to an excess of Type I errors (i.e., we reject the null hypothesis more often than the α level indicates).

A better approach is to replace each score by its absolute distance to the *median* of its group. Specifically, each score is replaced by

$$w_{as} = |Y_{as} - Md_a|,$$

where Md_a = median of Group a . This transformation gives very satisfactory results for an omnibus testing of the homogeneity of variance assumption. However, in order to implement more-sophisticated statistical procedures (e.g., contrast analyses, multiples comparisons), an even better transformation was proposed by O'Brien. Here, the scores are transformed as follows:

$$u_{as} = \frac{N_a(N_a - 1.5)(Y_{as} - M_a)^2 - .5SS_a}{(N_a - 1)(N_a - 2)},$$

where

N_a is the number of observations of Group a ,

M_a is the mean of Group a ,

SS_a is the sum of the squares of Group a :

$$SS_a = \sum_s (Y_{as} - M_a)^2.$$

When all the experimental groups have the same size, this formula can be simplified as

$$u_{as} = \frac{N(N - 1.5)(Y_{as} - M_a)^2 - .5SS_a}{(N - 1)(N - 2)},$$

where N = number of observations per group.

One Is a Bun . . .

To detail the computation of the median and the O'Brien transforms, we will use data from a memory experiment reported by Hunter. In this experiment, Hunter wanted to demonstrate that it is easier to remember an arbitrary list of words when we use a mnemonic device such as the peg-word technique. In this experiment, 64 participants were assigned to either the control or the experimental group. The task for all participants was to learn an arbitrary list of pairs of words such as "one-sugar," "two-tiger," . . . "ten-butterfly." Participants in the control group were told to try to remember the words as best as they could. Participants in the experimental group were given the following instructions: A good way to remember a list is to first learn a "nursery-rhyme" such as "*One is a bun, two is a shoe, three is a tree, four is a door, five is a hive, six is a stick, seven is heaven, eight is a gate, nine is a mine, and ten is a hen.*" When you need to learn a pair of words, start by making a mental image of the number, and then make a mental image of the second word and try to link these two images. For example, in order to learn "one-cigarette," imagine a cartoon-like bun smoking a cigarette.

Ten minutes after they had received their list, all the participants were asked to recall as many pairs as they could. The results are given in Table 1.

The results of this experiment are illustrated in Figure 1; they show that the participants from the experimental group did better than the participants from the control group did. To confirm this interpretation, an ANOVA was performed (see Table 2), and the F test indicates that, indeed, the average number of words recalled is significantly larger in the experimental group than in the control group.

Figure 1 also shows that a large proportion of the participants of the experimental group got a perfect score of 10 out of 10 words (cf. the peak at 10 for this group). This is called a *ceiling* effect: Some of the participants of the experimental group could have performed even better if they had had more words to learn. As a consequence of this ceiling effect, the variance of the experimental group is likely to be

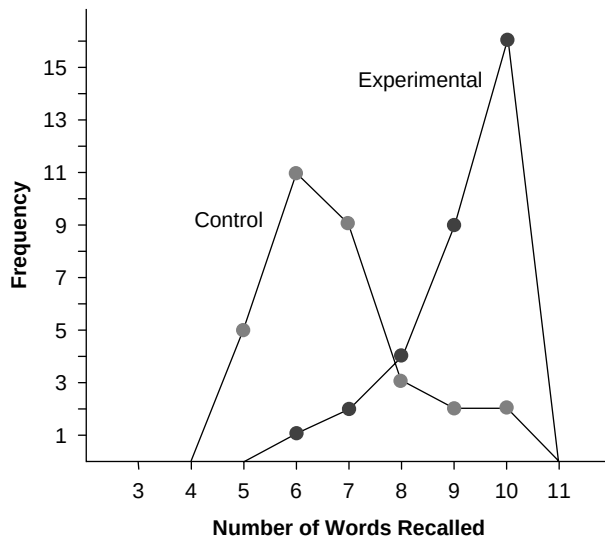
Table 1 Frequency of Subjects Recalling a Given Number of Words*

Number of Words Recalled	Control Group	Experimental Group
5	5	0
6	11	1
7	9	2
8	3	4
9	2	9
10	2	16
$Y_{a.}$	216	293
$M_{a.}$	6.750	9.156
$Md_{a.}$	6.500	9.500
SS_a	58.000	36.219

Source: Hunter (1964).

* For example, 11 participants in the control group recalled 6 words from the list they had learned.

smaller than it should be because the ceiling effect eliminates the differences between the participants with a perfect score. In order to decide whether this ceiling effect does in fact reduce the size of the variance of the experimental group, we need to compare the variance of the two groups.

**Figure 1** Results of the “Peg-Word” Experiment (One Is a Bun) From Hunter**Table 2** ANOVA Table for the Experiment of Hunter

Source	df	SS	MS	F	Pr(F)
Experimental	1	92.64	92.64	60.96**	.000000001
Error	62	94.22	1.52		
Total	63	186.86			

Source: Hunter (1964).

Note:** $p < \alpha = .01$. $R^2_{A,Y} = .496$.

The first step to test the homogeneity of variance is to transform the original scores. For example, the transformation of Y_{as} into w_{as} for Observation number 5 from the control group gives

$$w_{as} = |Y_{as} - Md_{a.}| = 5 - 6.5 = 1.5.$$

The transformation of Y_{as} into u_{as} for Observation number 5 from the control group gives

$$\begin{aligned} u_{as} &= \frac{N(N-1.5)(Y_{as} - M_{a.})^2 - .5SS_a}{(N-1)(N-2)} \\ &= \frac{32(32-1.5)(5-6.75)^2 - .5 \times 58}{31 \times 30} \\ &= 3.1828. \end{aligned}$$

The recoded scores are given in Tables 3 and 4. The ANOVA table obtained from the analysis of

Table 3 Recoded Scores: Control Group

Number of Words Recalled	Control Group		
	Frequency	w_{as}	u_{as}
5	5	1.5	3.1828
6	11	0.5	0.5591
7	9	0.5	0.0344
8	3	1.5	1.6086
9	2	2.5	5.2817
10	2	3.5	11.0538

Table 4 Recoded Scores: Experimental Group

Number of Words Recalled	Experimental Group		
	Frequency	w_{as}	u_{as}
5	0	4.5	
6	1	3.5	10.4352
7	2	2.5	4.8599
8	4	1.5	1.3836
9	9	0.5	0.0061
10	16	0.5	0.7277

Table 5 ANOVA Homogeneity of Variance Test (Recoded Scores: w_{as} , [Median])

Source	df	SS	MS	F	Pr(F)
Experimental	1	0.77	0.77	1.16 ^{ns}	.2857
Error	62	41.09	0.66		
Total	63	41.86			

Note: ns = no significant difference.

Table 6 ANOVA Homogeneity of Variance Test (Recoded Scores: u_{as} , [O'Brien Test])

Source	df	SS	MS	F	Pr(F)
Experimental	1	7.90	7.90	1.29 ^{ns}	.2595
Error	62	378.59	6.11		
Total	63	386.49			

Note: ns = no significant difference.

transformation w_{as} is given in Table 5. The ANOVA table obtained from the analysis of transformation u_{as} is given in Table 6.

A comparison of Tables 5 and 6 indicates that we cannot show that the ceiling effect observed in Figure 1 significantly reduces the variance of the experimental group compared to the control group.

—Hervé Abdi

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Homogeneity of Variance

Further Reading

- Abdi, H. (1987). *Introduction au traitement statistique des données expérimentales*. Grenoble, France: Presses Universitaires de Grenoble.
- Games, P. A., Keselman, H. J., & Clinch, J. J. (1979). Tests for homogeneity of variance in factorial designs. *Psychological Bulletin*, 86, 978–984.
- Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.
- Hunter, I. M. L. (1964). *Memory*. London: Penguin.
- Levene, H. (1960). Robust test for the equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press.
- Martin, C. G., & Games, P. A. (1977). ANOVA tests for homogeneity of variance: Non-normality and unequal samples. *Journal of Educational Statistics*, 2, 187–206.
- Miller, R. G. (1968). Jackknifing variances. *Annals of Mathematical Statistics*, 39, 567–582.
- O'Brien, R. G. (1979). A general ANOVA method for robust test of additive models for variance. *Journal of the American Statistical Association*, 74, 877–880.
- O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, 89(3), 570–574.

OBSERVATIONAL STUDIES

Both observational and experimental studies attempt to provide a valid estimate of the causal effect of some independent variable. Major differences exist, however, between these two types of investigation with respect to design, practicality, interpretation, and appropriate methods of analysis. For the sake of exposition, we will primarily consider two-group situations, although these two types of studies can involve comparisons of more than two groups. The principles and issues involved easily generalize to situations with more than two groups.

Design and Practicality Issues

Two key design issues distinguish observational and experimental studies: (a) the method used to form the comparison groups and (b) the nature of the process that determines whether a participant receives the treatment condition or the control condition. The

two-group experiment involves two essential characteristics: (a) random assignment of available participants to form two comparison groups and (b) experimenter-dictated assignment of the treatment condition to all participants in one group and the control condition to all participants in the other group. Because random assignment yields two groups that are probabilistically equivalent on all variables (before treatments are applied), a simple comparison of the group means (or medians) usually provides a meaningful estimate of the causal effect. In many situations, however, random assignment and experimenter-controlled treatment manipulation are impractical or unethical. For example, if one suspects that cocaine use by mothers causes brain damage in neonates, it would be unethical to randomly assign a sample of pregnant mothers to two groups, with one group receiving cocaine and the other (cocaine-free) group serving as the control group. Observational studies are often useful in such situations.

The observational study has neither random assignment of participants to form comparison groups nor experimenter-controlled assignment of treatment and control conditions to the groups. Rather, data are obtained from participants in nonrandomly formed groups: one that received the treatment condition and one that did not receive the treatment condition. Participants' exposure to the treatment occurs for some reason(s) other than the action of an experimenter. Suppose, for example, that two samples of mothers are identified: One sample used cocaine during pregnancy and the other did not. Although a comparison of the two samples on a measure of neonatal brain damage may seem interesting, this comparison does not necessarily provide an estimate of the causal effect. Such a comparison is often called the *naive effect estimate* because it almost certainly provides an invalid (biased) estimate of the causal effect.

Key Interpretational Concern

Naive estimates almost always contain bias because differences between treatment and control groups on the outcome may exist for many systematic reasons other than causal effects of treatments. Consequently, one does not know how much of this difference

should be attributed to initial group differences and how much (if any) should be attributed to treatment effects. It is naive to presume that the outcome difference reflects only treatment effects.

Analytic Issues

A thorough evaluation of an observational study will include statistical analyses that go beyond naive estimation to yield better estimates of the putative causal effect of the treatments. These analyses focus on two types of bias-producing variables: *overt*, which have been accurately measured on each participant before treatments were applied, and *hidden*, which have not been so measured.

Overt Bias Adjustment

The analysis of covariance (ANCOVA or the equivalent regression model), simple matched sampling, and subclassification are the traditional methods of adjusting for overt bias. Modern approaches represent extensions and refinements of these methods. One of the more useful modern approaches involves the *propensity score*, a number that indicates the propensity for a participant to fall in the treatment group rather than in the control group. Propensity scores can identify participants who are essentially equivalent on a very large number of variables. Once appropriate comparison participants are identified by means of propensity scores, the outcome analysis may take the form of a simple comparison by means of conventional parametric or nonparametric methods or may involve a more complex strategy that uses a modified form of ANCOVA. The latter approach often leads to a solution that is both less biased and more powerful than that of the simpler methods.

A classic example of observational research involved a large study of the effects of smoking. Three groups were compared: The average mortality rates were 13.5, 13.5, and 17.4 for nonsmokers, cigarette smokers, and cigar and pipe smokers, respectively. A comparison of these rates (i.e., the naive estimates) suggested that there was no effect of cigarette smoking relative to not smoking and that cigar and pipe smoking increased mortality. After adjusting the

mortality rates for age, however, these rates were 13.5, 21.2, and 13.7 for nonsmokers, cigarette smokers, and cigar and pipe smokers, respectively. Hence, after adjusting for overt bias associated with age differences among groups, cigarette smoking appeared to increase mortality.

Hidden Bias Adjustments

Although modern methods of analysis provide satisfactory solutions to the problem of overt bias, one always has to acknowledge that unknown but important (i.e., bias-causing) variables probably have been left out of the overt bias adjustments. For example, Sir Ronald Fisher seriously entertained the notion that there may be genetic differences between smokers and nonsmokers before they begin to smoke. If unmeasured genetic characteristics have a causal influence on the between-group difference on mortality rates, they qualify as sources of hidden bias.

Because the exclusion of such variables can completely change the size and sign of the treatment-effect estimate, a statement acknowledging this possibility should be included in every document reporting the results of an observational study. It is often possible, however, to apply additional methods to examine the plausibility of the hidden variable explanation for the results. Consequently, many observational studies rely on assembling information outside the group comparison to bolster causal conclusions. This additional information often takes the form of (a) a second control group, (b) additional results that are consistent with a theory, and (c) sensitivity analysis.

Suppose an observational study of two methods of teaching reading has found an important difference between the treatment and control groups on the outcome reading measure. Further, suppose that the variable of age, which was not included in the original analysis (and can no longer be obtained), is a source of hidden bias. In this case, it would be reasonable to identify an additional group (the second control), believed to differ in age relative to the original control group. A comparison is then made between the two control groups on the outcome measure; if the two

differ on the reading outcome in the suggested direction (i.e., the control group with the higher average age also has the higher reading score), this information provides support for the hypothesis that age is a hidden bias in the comparison of the treatment and control groups. In this case, the original interpretation of a differential treatment effect is in doubt.

One can often use a theory to predict a pattern of differences between the means of certain groups. For example, suppose that the reading study mentioned above involved a direct instruction group and a conventional reading (control) group. Perhaps the amount of time a person is exposed to a specific aspect of direct reading instruction is the critical component of the causal mechanism, and the observational study supports the superiority of this treatment. If existing evaluative data comparing three other reading programs are available, and if each of these programs differs on a measured amount of the critical component, a pattern of results can be predicted from this information. If the pattern reveals a clear association between the amount of critical component of the direct reading method and the effectiveness of the three treatments, the claimed treatment effect in the observational study is supported. Alternatively, if both the difference in the original observational study and the differences found among the means in the supplemental data are associated with age, the hidden bias explanation is supported.

The third approach for considering hidden bias requires no additional data. Instead, the original observational study data are subjected to what is known as a *sensitivity analysis*. This type of analysis evaluates how large a potential hidden bias would have to be to produce meaningful changes in the results of the observational study. Some versions of sensitivity analysis produce bounds on *p* values that are associated with a specified size of hidden bias. Other methods provide a measure of how much two groups must differ on a hidden variable for the results to change from statistically significant to nonsignificant.

—Bradley E. Huitema
and Sean Laraway

See also Authenticity; Interrater Reliability

Further Reading

Huitema, B. E. (in preparation). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and observational studies* (2nd ed.). Hoboken, NJ: Wiley.

Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer-Verlag.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.

Propensity score reference list (annotated) with examples of observational studies: <http://sswnt5.sowo.unc.edu/VRC/Lectures/>

OCKHAM'S RAZOR

Ockham's razor is a methodological principle introduced by the medieval Franciscan friar and philosopher-theologian William of Ockham. He was born in 1285 in the village of Ockham, in Sussex, England, and died probably around 1347, during the plague epidemic in Munich, where he had spent his latter years at the court of the outlawed German-Roman emperor Ludwig of Bavaria as one of his theological-political advisers.

The razor principle was stated by Ockham several times in his writings. It is introduced as a kind of theorem or axiom that is never proven but taken as evident and hence not in need of further proof. In a well-known version, it reads “pluralitas non est ponenda sine necessitate—a plurality [of entities] should not be postulated without necessity.” To understand what Ockham meant by that, it is necessary to introduce very briefly the theological and philosophical background discussions during his time, as well as key aspects of his life.

Historical Background

At the beginning of the 13th century, two new spiritual movements arose within a very short period of two decades and grew to a powerful position in medieval society: the so-called mendicant friars of the Franciscan and Dominican orders. The key elements of

the Dominican calling were preaching to Christians and heretics alike and most important, converting all who were considered “heathens” (i.e., people of Jewish, Muslim, or nonorthodox Christian faiths) and establishing the knowledge, background, and learning to be able to do so. Therefore, Dominicans pressed into the newly founded universities in Paris, Oxford, and Cambridge to install a pertinent training, teaching, and research base for their enterprise. They were soon followed by the Franciscan friars. Although learning and academic work had not been a primary aspect of early Franciscan spirituality, Franciscans, too, soon understood that the Franciscan order also had to train its new members well if they were to complete their mission of reinvigorating Christian life and service to the community. Thus, the academic life and debates in the 13th century were largely dominated by these new movements and their academic beacons at the universities. One major challenge for these academics was to integrate Aristotelian science and philosophy into the traditional Christian teaching. Knowledge about Aristotle's philosophy had been scarce and based on only a few writings known to the West. Through both peaceful contacts with Islamic science, mainly in Spain and Sicily, and the combats surrounding the Crusades, such as the crusaders' sacking of Constantinople, more Aristotelian writings came to the West, and as a consequence, Aristotelian teaching became known. Thomas Aquinas (1225–1274), the mastermind of the Dominican order at that time, clearly saw that only a Christian reinterpretation of Aristotle would save a coherent picture of the world and give enough rational appeal to Christian theology, both for the satisfaction of critical Christian thinkers and philosophers and for armament against the Muslim theologians who had already integrated Aristotle into their worldview. Thus, Thomas Aquinas devoted his academic career to interpreting and amending Aristotle and reconciling him with Christian teachings. Part of this enterprise was the adoption of Aristotle's theory of perception and mental activity as expounded in his book *De Anima* [*On the Soul*]. The theory, briefly, supposes that in visual perception, for instance, the eyes extract what in Latin was called a *species sensibilis*, a perceptual template, if we may interpret freely. This

template is presented to the mind, which from it abstracts another template, which in this case is mental. This second template is the basis for further operations. Thus, a whole series of intermediaries were introduced that would mediate between the object of perception and the mental operations of the mind.

The Franciscan calling was, in general, different from that of the Dominicans, and a latent competition ran through all the two orders' activities. Franciscans put more emphasis on personal piety, inner experience, and the fostering of a religious life in general. One group in particular, whose members called themselves "Spirituals," tried to go back to what was seen as Saint Francis's original testament: a poor life as a mendicant order, little interaction with the authorities of church or world, and a support of private piety through prayer and contemplation, which should lead each Christian to an intimate personal relationship with Christ.

When William of Ockham entered the Franciscan order, presumably around the age of 14, and his scholarly training started, this debate was reaching a first climax, and it can be safely assumed that he was drawn into it. He was sent to Paris by his order, as every intelligent and promising young student would have been in those days, to finish his studies. When he arrived there, the debates around the Aristotelian-Thomastic theory of perception were at their height. Ockham, however, had had other training: One of his teachers, Roger Bacon, another Franciscan, advocated experiential knowledge and in particular inner experience as an important way to know both the world and God's will. Experience always had an element of immediacy, of unmediated access to its object. Now, if you were to apply the then-current general Aristotelian theory of knowledge to direct experiences, say of an inner mystical relationship with God, what would that entail? You would have to abstract from your own inner experience a species, which would again be abstracted by another inner act until it reached your mind as a diluted and colorless mental template of what originally was a rich experience. Even with very simple personal experiences, such as hunger, thirst, or pleasure, the Aristotelian concept of perception would distort experience by abstractions. And thus, a systematic philosophical and epistemological

place for direct experience could not be had as long as this philosophy of perception was the generally accepted one.

This state of the debate was certainly one motive for Ockham's critique, for he had understood a very important precondition: If there should ever be a systematic place for a direct inner (and outer) experience of mental events in any epistemology, then the Scholastic concept of perception introduced by Thomas Aquinas must be critiqued. If there were to be any possibility of a direct inner contact of a soul with God, as Franciscan piety was striving for, then the Scholastic theory of perception must be replaced by something simpler. Thus, one very powerful motive, among others, for Ockham was the wish to establish a place for direct inner experience within epistemology and for experience in general as an access both to the world and to one's inner being. Thus, as a theorem from which to start his critiques, he introduced his razor principle: No entities beyond necessity.

Consequences and Place Within the Scientific Enterprise

This principle of parsimony, as it is also called, is intuitively appealing. It was introduced as a guideline that should inform our thinking and theorizing. Ockham used it to deconstruct the Thomistic and Scholastic theory of perception and to find a place for experience and immediate access to reality.

Another important debate of the time pitted the theory of universals against nominalism. While a universalist can sit in a study thinking about universals and their relationship with the world, a nominalist eventually has to go outside and study all individual specimens of, say, horses to understand what it is to be a horse. Ockham's razor, applied to this debate, led directly to an empiricist approach to reality. Thus, Ockham's endeavor, philosophical and abstract as it might seem today, was the very precondition for modern science as an empirical enterprise.

This might be the reason that the principle of parsimony is held dear by modern science. Whenever someone announces the discovery of something new—a new species of animal, energy, or any other

entity—science as a body reacts skeptically, sharpening Ockham's razor. (In this way, Ockham's razor is perhaps also a way to keep jerks and quacks out of science.) It is at the base of science's inherently reductionist approach: Whenever some seemingly unknown phenomenon is discovered or whenever something is not well understood, the historically most promising approach is not to postulate a new entity but to try to understand it through already known principles and reduce the phenomenon to a set of theories or rules that are already understood.

An Example

A well-known historical example is the explanation of mesmerism as hypnotic effects. In the 18th century, the German physician Franz Anton Mesmer had postulated a new type of energy, which he called "animal magnetism." He thought that this energy would emanate from every living being, with humans being especially endowed with it, and that in illness, this energy was disturbed. Consequently, he used his own "animal magnetism" to heal sick patients. He was a center of attention in pre-Revolutionary Paris, with poor people and the aristocracy alike using his healing powers, apparently quite successfully. Mesmer's claims were scrutinized by the French Academy of Sciences. When shielded by a curtain, the purported magnetic phenomena did not occur, but they were clearly evident provided the healer was able to stay in visual contact with his patient. The Academy concluded that animal magnetism had not been proven and that the phenomena were likely due to suggestion and hypnotism. In other words, Ockham's razor was implicitly used to "cut out" the newly introduced notion of animal magnetism and reduce it to known phenomena: suggestion and hypnotism.

Thus, the razor principle introduces a sound element of skepticism and in fact conservatism into the scientific debate, although it had been used by its inventor as a weapon against conservative mainstream theorizing and as a motor of progressive methodology. It is quite fitting that William Ockham was called *inceptor venerabilis* (venerable initiator) in his day. Nowadays, the razor principle has turned from a

sword of the vanguard to a shield of the establishment. Now Ockham's razor has to be balanced with a stance that is true to what we see and experience.

Thus, Ockham's razor, the methodological principle of parsimony, has to be seen in context. Its function is to prevent scientists and the scientific community at large from leaping to accept new explanations for phenomena before it is clear that the phenomena cannot be explained by existing models. Only if these models fail can new categories of phenomena, new theories, or a new worldview be considered.

Ockham's razor is the first methodological principle introduced by a medieval scholar without an apparent source in earlier or antique writings. It marks the beginning of a modern era of potentially unprejudiced investigation and discovery and has served scholars and scientists well for more than 700 years. It is part and parcel of the scientific enterprise. In order not to be perverted into an instrument of hegemony of one doctrine, it needs balancing by a generic stance of openness toward phenomena, a stance some of us refer to as Plato's Life Boat.

—Harald Walach

See also Chance; Significance Level; Type I Error

Further Reading

- Böhner, P. (1943). The notitia intuitiva of nonexistents according to William Ockham. *Traditio*, 1, 223–275.
- Day, S. J. (1947). *Intuitive cognition: A key to the significance of the late Scholastics*. St. Bonaventure, NY: Franciscan Institute.
- Leff, G. (1975). *William of Ockham: The metamorphosis of Scholastic discourse*. Manchester, UK: Manchester University Press.
- Leppin, V. (2003). *Wilhelm von Ockham: Gelehrter, Streiter, Bettelmönch*. Darmstadt, Germany: Wissenschaftliche Buchgesellschaft.
- McCord Adams, M. (1970). Intuitive cognition, certainty, and scepticism in William Ockham. *Traditio*, 26, 389–398.
- McCord Adams, M. (1987). *William Ockham* (Vols. 1–2). Notre Dame, IN: University of Notre Dame Press.
- Ockham, W. v. (1982). *Quaestiones in III librum sententiarum*. In F. E. Kelley & G. I. Etzkorn (Eds.), *Opera theologica: Vol. 6* (pp. 43–97). St. Bonaventure, NY: Franciscan Institute.

Walach, H., & Schmidt, S. (2005). Repairing Plato's life boat with Ockham's razor: The important function of research in anomalies for mainstream science. *Journal of Consciousness Studies*, 12(2), 52–70.

OGIVE

An ogive is a line plot of the cumulative frequency distribution against values of the random variable. Francis Galton coined the term *ogive* to describe the shape of the normal cumulative distribution function, as it has a form similar to the S-shaped Gothic ogival arch.

Table 1 Student Test Scores

Test Scores	0–19	20–39	40–59	60–79	80–100
Frequency	1	6	22	17	4
Relative frequency	0.02	0.12	0.44	0.34	0.08
Cumulative frequency	0.02	0.14	0.58	0.92	1.00

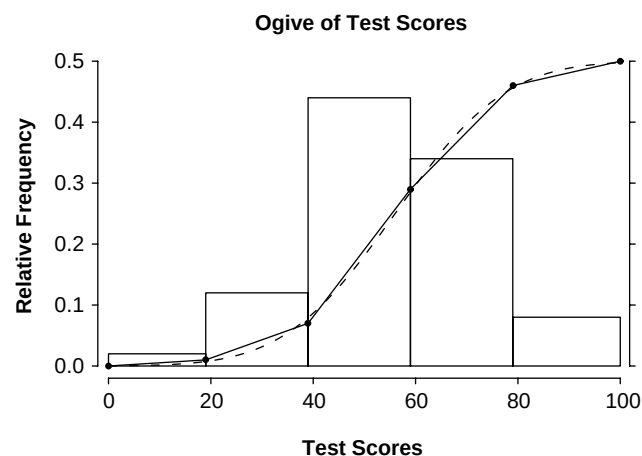


Figure 1 Ogive Plot of Student Test Scores With Frequency Distribution and Normal Cumulative Distribution Curve

Note: Solid line is ogive plot of student test scores with frequency distribution, and dashed line is normal cumulative distribution curve

The ogive can display the population cumulative frequency distribution or an estimate from a sample. The random variable can be continuous or discrete (as long as there is a natural ordering of the outcomes). The range of possible outcomes is divided into classes. A line is drawn to connect points at the upper limit of each class interval and the cumulative frequency distribution. For a discrete random variable, each possible outcome can be a class, and therefore the upper limit of the class is the outcome itself.

The student test scores in Table 1 are split up into five class intervals. The ogive is often overlaid with the frequency distribution, as in Figure 1. The solid line is the empirical cumulative distribution, and the

Table 2 Frequency Distribution of a Discrete Data Set With Number of Minor Accidents in a Factory per Year

Number of Accidents	0	1	2	3	4	5	6
Frequency	3	5	6	3	2	1	0
Relative frequency	0.15	0.25	0.30	0.15	0.10	0.05	0
Cumulative frequency	0.15	0.40	0.70	0.85	0.95	1	1

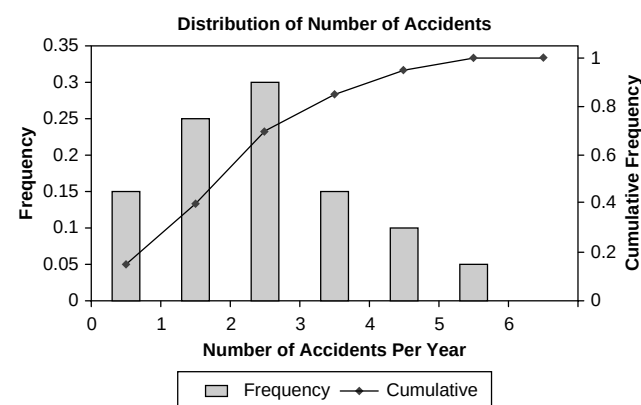


Figure 2 Ogive Plot of Number of Accidents in a Factory per Year With Frequency Distribution

dashed line is an estimate of the population distribution assuming normality (notice the distinctive S shape observed by Galton).

Table 2 summarizes the frequency distribution of a discrete data set consisting of the number of minor accidents recorded in a factory per year. The corresponding ogive is shown in Figure 2, which was created with Excel. Notice that the ogive points are now at the outcome values rather than at the upper limit, which was the case in the previous data set.

—Carl J. Scarrott

See also Cumulative Frequency Distribution; Frequency Distribution

Further Reading

Kenney, J. F., & Keeping, E. S. (1962). *Mathematics of statistics: Part I* (3rd ed.). Princeton, NJ: Van Nostrand.

Ogive (history and definition): <http://www.pballew.net/arithmetic.html#ogive>

ONE- AND TWO-TAILED TESTS

The one-tailed or two-tailed test is a part of a much more elaborate procedure called hypothesis testing or tests of statistical significance. Prior to doing a hypothesis test, a researcher will have formulated a problem, identified the variables, formulated the hypotheses, and collected the data.

In hypothesis testing, there are five basic steps:

- Stating the null hypothesis
- Stating the alternative hypothesis
- Computing the test statistic
- Formulating the decision rule and making a decision
- Drawing a conclusion

At the very basic level, the researcher would be comparing either a sample mean against a population mean or the difference between two sample means. Hypothesis tests are not restricted to tests on means. Other comparisons and evaluations could involve

variances, proportions, and correlations, to name a few. Furthermore, these tests are not restricted to comparing only two means.

The statement of the alternative hypothesis is a statement expressed in statistical terms concerning an investigator's research interest. For example, the investigator may have a question concerning the effects of alcohol consumption on perceptual judgment. If the investigator feels from personal experience or observation that alcohol would have a negative effect on perceptual ability, the alternative hypothesis would reflect that. For research problems in which the data are considered parametric, the null and alternative hypotheses would be expressed by the use of population parameters. The simplest case would be between two means, μ_1 and μ_2 .

In the example, alcohol is given to an experimental group of participants, μ_1 and a placebo, an inactive simulation of alcohol, is given to a control group, μ_2 . Each participant is measured on the number of correct judgments made on a perceptual task. The alternative hypothesis would state that the group receiving alcohol (the experimental group) would perform worse than the group receiving the placebo (the control group). Statistically, the alternative hypothesis would be written as $H_1: \mu_1 - \mu_2 < 0$ or $H_1: \mu_1 < \mu_2$. The alternative hypothesis is not directly testable. Note that it just says that one group is worse than the other. It does not say by how much. To test this hypothesis directly would require a large, perhaps infinitely large, number of tests in which each difference was to be tested. It is much easier to create a hypothesis that is the opposite of the alternative hypothesis and, with empirical evidence, demonstrate that it cannot be tenable. If this opposite hypothesis is not tenable, then by inference, the alternative hypothesis must be true. This opposite hypothesis is called the null hypothesis. Some part of the null hypothesis points directly to a testable value, such as zero. So for this example, the null hypothesis, written as H_0 , is $H_0: \mu_1 - \mu_2 \geq 0$. Note that the null hypothesis contains the equal sign.

The null hypothesis is used to help direct the hypothesis test. The investigator would assume the null hypothesis to be true and then, through empirical data, demonstrate that it cannot be tenable. As a result

of this, the alternative hypothesis is shown to be tenable. However, it is the alternative hypothesis that dictates whether the test against the null hypothesis should be one-sided or two-sided. Whenever the investigator's research question indicates that one treatment group is better, more improved, more impaired, weaker, faster, or something else along these lines, than another group, the alternative hypothesis would be considered as one-tailed, or one-sided. When no direction is given by the investigator as to which group will be better or worse than the other group, the alternative hypothesis would be two-tailed. Consider the example given. The investigator had hypothesized that alcohol consumption would lead to impaired performance compared to the performance of those who did not consume alcohol. The alternative hypothesis was $H_1: \mu_1 < \mu_2$. This is a one-tailed test. If the investigator had hypothesized that those that consumed alcohol would demonstrate improved performance over those with no alcohol, the alternative hypothesis would have been written as $H_1: \mu_1 > \mu_2$. This still would have been a one-tailed test. However, if the investigator had stated uncertainty as to whether or not alcohol consumption would lead to a positive or negative change in performance in comparison to those that did not consume alcohol, the test would be two-tailed. It is two-tailed because the uncertainty would make it possible that either $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$ could be true if the null hypothesis were shown not to be tenable. For this situation, the null hypothesis would have been stated as $H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$. Such a test is also called a *nondirectional* test.

A *two-tailed*, or *two-sided*, test is so named because if the normal distribution, or the t distribution, is used, the two tails, or the two sides, of the distribution are employed in the estimation of probabilities. Consider a .05 probability of a Type I error. With a normal-shaped sampling distribution, .025 of the area of the curve falls to the right of 1.96 standard deviation units above the mean, and .025 falls to the left of 1.96 standard deviation units below the mean (See Figure 1). The sum of these areas is 5% of the total area under the curve ($\alpha = .05$). In a hypothesis test, the chances are 2.5 in 100 of getting a difference of 1.96 standard deviation units in one direction due to

chance factors alone, and 2.5 in 100 in the other direction. Hence the total by chance alone in both directions is 5 in 100. For significance at the 1% level, a value of 2.575 is required for a nondirectional test. For one-tailed tests, if the normal, or t , distribution is used, only one of the tails is used to estimate the required probabilities. For a Type I error probability of .05, the entire amount is placed into one of the tails of the distribution. For a right-tailed, one-tailed test, the required critical value needed to reject the null hypothesis is +1.645 (see Figure 2). For a left-tailed test, the critical value is -1.645 (see Figure 3). With one-tailed tests, the investigator could make an error in choosing the wrong direction or tail contrary to the statement made by the alternative hypothesis.

In summary, for a one-tailed case, an investigator has only one critical value to contend with when doing the hypothesis test, and the two-tailed test has two critical values. For a fixed probability of a Type I

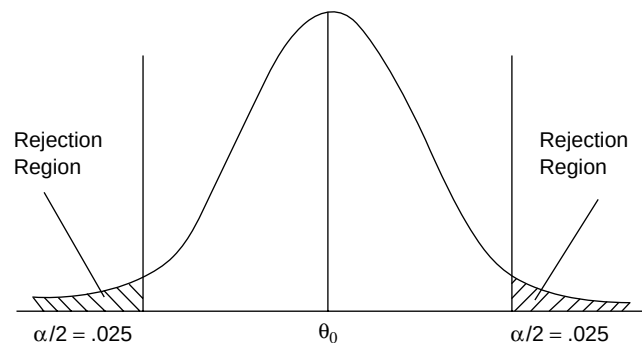


Figure 1 A Two-Tailed Test of the H_0 With $\alpha = .05$

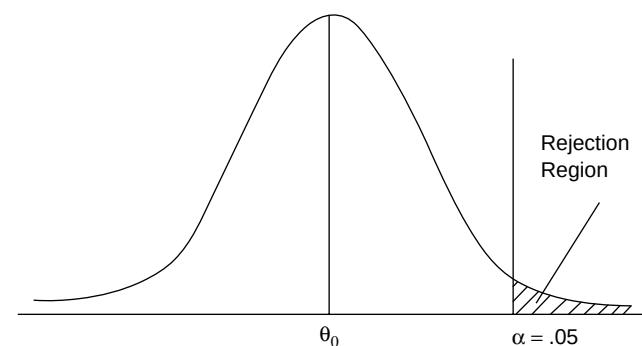


Figure 2 A One-Tailed Test (*Right*) of the H_0 With $\alpha = .05$

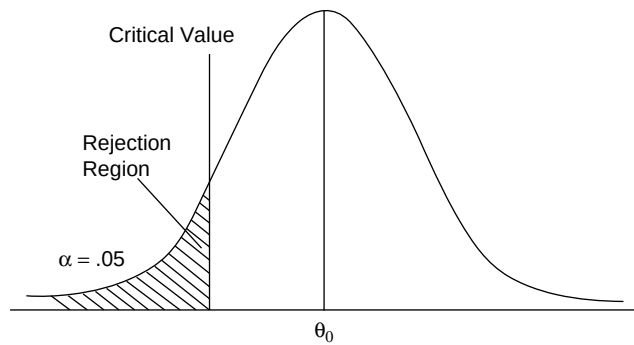


Figure 3 A One-Tailed Test (Left) of the H_0 With $\alpha = .05$

error, α , the chances of rejecting the null hypothesis are higher for a one-tailed test than for a two-tailed test. The reason is that the α value is divided in half and placed across two tails of distribution in the two-tailed situation whereas in the one-tailed case, the entire amount is spread over only one of the tails of the distribution. In terms of statistical power, the one-tailed test is more powerful than the two-tailed test provided the investigator has hypothesized the proper direction of the statistical test.

In the social and behavioral sciences, there has been much discussion concerning the use of one- and two-tailed tests. One of the major issues these discussions have centered on is an investigator's decision to change from an originally stated two-tailed test to a one-tailed test after the hypothesis test has been done. The two-tailed test yielded a nonsignificant result, but on reflection, the investigator decided a one-sided test would have produced a significant result. Some have claimed that the decision to choose the alternative and null hypotheses post hoc is unethical. Plus, investigators should always remind themselves that significant results could still be due to chance.

Another issue is what an investigator should do if a two-tailed test turns out significant. With a significant nondirectional test, the investigator knows only that the two groups are different. A very stringently agreed on nondirectional test does not allow the investigator to make statements of a negative or positive effect when inspecting the group means. Some statisticians have proposed a "three-decision procedure." In this procedure, the investigator has three possibilities:

- Do not reject H_0 .
- Reject H_0 in favor of alternatives on one side.
- Reject H_0 in favor of alternatives on the other side.

In this three-decision procedure, however, there are six kinds of possible errors instead of two, with the probabilities of the errors varying. In some cases, the probabilities are small enough to ignore. Some have considered this three-decision procedure to be a composite of two one-tailed tests.

—Howard B. Lee

Further Reading

Comrey, A. L., & Lee, H. B. (1995). *Elementary statistics: A problem solving approach* (3rd ed.). Dubuque, IA: Kendall-Hunt.

Statistical significance discussion: <http://www.statpac.com/surveys/statistical-significance.htm>

ONE-WAY ANALYSIS OF VARIANCE

One-way analysis of variance is part of the family of tests known as analysis of variance (ANOVA). Typically, it is used to analyze experimental designs in which only one independent variable has been manipulated. Usually, one-way ANOVA is used to test whether differences exist between three or more means; however, it can be applied to situations in which there are only two means to be compared. Although the t test is preferred by many in such situations, the F test produced in one-way ANOVA is a direct function of t , and so it also is a legitimate way to compare two means.

The two types of one-way ANOVA differ in terms of the experimental design to which they are applied. If data representing different levels of an independent variable are independent (i.e., collected from different entities), then a *one-way independent* ANOVA (also called a *between-groups* ANOVA) can be used. When data are related, such as when different entities have provided data for all levels of an independent variable, a *one-way repeated measures* ANOVA (also

called a *within-subjects* ANOVA) can be employed. In both cases, the underlying principal is the same: A test statistic F is calculated that is the ratio of systematic variance (variance explained by the independent variable, that is, the experimental manipulation) to unsystematic variance (variance that cannot be explained, or error). If the observed value exceeds the critical value for a small probability (typically .05), we tend to infer that the model is a significant fit of the observed data or, in the case of experiments, that the experimental manipulation has had a significant effect on performance.

An Example

One study looked at the processes underlying obsessive-compulsive disorder by inducing a negative mood, a positive mood, or no mood in people and then asking them to imagine they were going on holiday and to generate as many things as they could that they should check before they left. The data are in Table 1. Three different groups of people participated in this experiment, each group representing a level of the independent variable, mood: negative mood, positive mood, and no mood induced. These data are independent because they came from different people. The dependent variable was the number of items that needed to be checked.

The starting point for ANOVA is to discover how much variability there is in the observed data. To do this, the difference between each observed data point and the grand mean is calculated. These values are then squared and added together to give us the total sum of squared error (SS_T):

$$SS_T = \sum_{i=1}^n (x_i - \bar{x}_{grand})^2.$$

Alternatively, this value can be calculated from the variance of all observations (the *grand variance*) by multiplying it by the sample size minus 1:

$$SS_T = s^2 (N - 1).$$

Table 1 Numbers of Things People in Negative or Positive Mood or No Induced Mood Thought They Should Check Before Going on Holiday

	Negative Mood	Positive Mood	No Mood Induced
	7	9	8
	5	12	5
	16	7	11
	13	3	9
	13	10	11
	24	4	10
	20	5	11
	10	4	10
	11	7	7
	7	9	5
$\bar{x}$	12.60	7.00	8.70
S^2	36.27	8.89	5.57
Grand Mean = 9.43		Grand Variance = 21.43	

Source: Davey et al. (2003).

Note: $\bar{x}$ = mean; S^2 = variance.

The degrees of freedom for this value are $N - 1$, where N is the total number of observations (in this case, 30). For these data, the total degrees of freedom are 29. For the data in Table 1, we get

$$SS_T = s_{grand}^2 (N - 1) = 21.42(30 - 1) = 621.18.$$

Having established the total variance to be explained, this variance is partitioned into two parts: variance explained by the linear model fitted to the data (expressed as the model sum of squared error, SS_M) and variance that the model cannot explain (expressed as the residual sum of squared error, SS_R). The model sum of squares is the squared difference between the grand mean and the values predicted by the linear model. When analyzing data from groups (i.e., when an experiment has been conducted), the linear model takes the form of the group means. The variance explained by the model is, therefore,

$$SS_M = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_{grand})^2,$$

in which k is the number of groups. For the data in Table 1, we would get

$$SS_M = 10(12.60 - 9.43)^2 + 10(7.00 - 9.43)^2 + 10(8.70 - 9.43)^2 = 164.87.$$

The degrees of freedom (df_M) for this sum of squared error are 1 less than the number of parameters estimated: When data from groups are analyzed, this is the number of groups minus 1 (denoted as $k - 1$), in this case 2, and when continuous predictors are used, it is the number of predictors.

Finally, we need to establish the error in the model, or the variance not explained by the model. This is simply the squared difference between the values predicted by the model and the observed values. When analyzing data from groups, the model used is the group means, so we are looking at the squared difference between the observed value and the mean of the group from which that observation came:

$$SS_R = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2.$$

This can be simplified by using the variance within each group (S_j^2) and reexpressing the equation as

$$SS_R = \sum_{j=1}^k s_j^2 (n_j - 1).$$

For the data in Table 1, this gives us

$$\begin{aligned} SS_R &= 36.27(10 - 1) + 8.89(10 - 1) + 5.57(10 - 1) \\ &= 326.43 + 80.01 + 50.13 \\ &= 456.57. \end{aligned}$$

Alternatively, the value can be derived from $SS_R = SS_T - SS_M$ (which gives the same answer to within

rounding error). The degrees of freedom for SS_R (df_R) is the total degrees of freedom minus the degrees of freedom for the model ($df_R = df_T - df_M = 29 - 2 = 27$). Put another way, it is $N - k$ (the total sample size, N , minus the number of groups, k).

SS_M tells us how much variation the model (e.g., the experimental manipulation) explains, and SS_R tells us how much variation is due to extraneous factors. However, because both of these values are summed values, they will be influenced by the number of scores that were summed. To eliminate this bias, we can calculate the average sum of squared error (known as the mean squared error, MS), which is simply the sum of squares divided by the degrees of freedom:

$$MS_M = \frac{SS_M}{df_M} = \frac{164.87}{2} = 82.44$$

$$MS_R = \frac{SS_R}{df_R} = \frac{456.57}{27} = 16.90$$

MS_M represents the average amount of variation explained by the model (e.g., the systematic variation attributable to the experimental manipulation), whereas MS_R is a gauge of the average amount of variation explained by extraneous variables (the unsystematic variation).

The F ratio is a measure of the ratio of the variation explained by the model and the variation explained by unsystematic factors. It can be calculated by dividing the model mean squared error by the residual mean squared error:

$$F = \frac{MS_M}{MS_R} = \frac{82.44}{16.90} = 65.54.$$

The F ratio is, therefore, a measure of the ratio of systematic variation to unsystematic variation, or a comparison of how good the model is compared to how bad it is. In experimental scenarios, it is the ratio of the experimental effect to the individual differences in performance. The observed value of F is compared to critical values of F from a special distribution known as the F distribution with df_M and df_R degrees of freedom, which represents the values of F that can

be expected at certain levels of probability. If this value is significant, then the model is a good fit of the data, or in experimental designs, the independent variable is believed to have had an effect on the dependent variable.

Following Up One-Way ANOVA

ANOVA is an omnibus test and so tells us only that the means of the groups differ from one another. When more than two means have been compared, the *F* test indicates only whether two or more of those means differ. Therefore, one-way ANOVA is usually followed up either with planned comparisons of specific sets of means or with post hoc tests, which compare all combinations of pairs of means (see, for example, the entries on Bonferroni Test and Newman-Keuls Test).

Assumptions

For the *F* ratio to be accurate, the following assumptions must be met: (a) Observations should be statistically independent, (b) data should be randomly sampled from the population of interest and measured at an interval level, (c) the outcome variable should be sampled from a normal distribution, and (d) there must be homogeneity of variance.

Differences With Repeated Measures

When data are related (i.e., the independent variable has been manipulated by use of the same entities), the basic logic described above still holds true. The resulting *F* can be interpreted in the same way, although the partitioning of variance differs somewhat. However, when a repeated measures design is used, the assumption of independence is violated, and this gives rise to an additional assumption of *sphericity*. This assumption requires that the variances of difference scores between conditions be roughly equal. When this assumption is not met, the degrees of freedom associated with the *F* value must be corrected with one of two estimates of sphericity: the Greenhouse-Geisser estimate or the Huynh-Feldt estimate.

—Andy P. Field

See also Analysis of Variance (ANOVA); Bonferroni Test; Dependent Variable; Homogeneity of Variance; Independent Variable; Newman-Keuls Test; Normal Curve

Further Reading

- Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). Perseveration of checking thoughts and mood-as-input hypothesis. *Journal of Behavior Therapy & Experimental Psychiatry*, 34, 141–160.
- Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury.

ORDINAL LEVEL OF MEASUREMENT

The ordinal level of measurement refers to measurement in which the distances between observed values are irrelevant and only the order relations of $<$, $>$, and $=$ should be considered.

Consider the Likert-type item common in the social sciences. Typically, such an item consists of a statement that captures the essence of the construct being measured (e.g., depression), followed by adjectives or adjectival response phrases indicating degree of endorsement of the statement. Very often, the adjectival phrases are coded by increasing integer values, an example being 0 = *not at all*, 1 = *a little bit*, 2 = *moderately*, 3 = *a lot*, 4 = *quite a bit*. It is important to realize that considered in isolation (e.g., not in the context of a scaling model), the statements are inexact to the extent that they do not imply distances between them. Equidistant interpretations such as *quite a bit* – *a lot* = *moderately* – *a little bit*, characteristic of interval-level measurement, are not justified. On the other hand, ordinal interpretations such as *not at all* $<$ *quite a bit* are justified. Suppose there is no response error for an item and Respondent A selects *not at all* and Respondent B selects *quite a bit*. Then the empirical ordering with respect to endorsement is Respondent A $<$ Respondent B, which is represented by $0 < 4$. The empirical ordering of the respondents is accurately represented by the order relations of the observed values. The representation is limited to the relations

of $<$, $>$, and $=$, because $4 - 3 = 2 - 1$, for example, is not a valid representation for ordinal level of measurement.

An interesting characteristic of the ordinal level of measurement is invariance under any order-preserving transformation such as the natural log transformation. Suppose we obtain the following observed values from five people using our Likert-type item: 1, 2, 2, 3, 4. Ordinal level of measurement dictates that the values of the number are irrelevant and only their order is important. Thus, any other set of numbers with the same order will be an equally valid ordinal representation. For example, taking the natural log of the original values, we obtain 0, .6932, .6932, 1.0986, 1.3863. The log-transformed values look quite different from the original numbers but retain the same ordinal information.

Depending on an analyst's epistemology, the monotonic invariance property may proscribe certain statistical methods for ordinal-level data. Methods such as the Spearman rank correlation coefficient yield invariant results under any monotonic transformation whereas methods such as Pearson's correlation coefficient do not. When one is analyzing ordinal data, it may be desirable to use a statistical method that considers only ordinal relations so that test statistics and p values do not change under order-preserving transformations.

—Jeffrey D. Long

See also Interval Level of Measurement; Nominal Level of Measurement; Ratio Level of Measurement

Further Reading

- Hand, D. M. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 159, 445–492.
- Michell, J. M. (1990). *An introduction to the logic of psychological measurement*. Mahwah, NJ: Erlbaum.
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

ORTHOGONAL PREDICTORS IN REGRESSION

In multiple regression, a single dependent variable (y , the criterion) is predicted from a number of

independent variables. If there were only a single independent variable (x_1), the regression coefficient b is simply a rescaling of the correlation coefficient r_{xy} (i.e., $b = r_{xy}$). Therefore, r_{xy} can be used to gauge the importance of the single predictor variable. With multiple predictor variables, the correlations among these predictors (i.e., multicollinearity), as well as the correlations of the predictors with the criterion, influence the regression weights. If one is interested only in optimizing the overall fit of the model (i.e., R^2), then the regression coefficients do not require much interpretation. However, understanding the effect of each individual variable is very difficult. It can also be the case that variables significantly correlated with the criterion do not enter the model because another variable is more strongly associated with the criterion and the additional variance explained is insufficient to achieve statistical significance.

One approach to this problem is to conduct a principal component analysis. All dimensions are independent but, unfortunately, hard to interpret with respect to the original variables. Alternatively, the predictor variables can be orthogonalized by means of a transformation so that all predictor variables are uncorrelated with each other, but simultaneously each transformed predictor variable is maximally correlated with the corresponding original variable. If the correlations of the transformed variables with the original predictor variables are high, the results of the orthogonal regression analysis are easy to interpret. The transformation provides variables that are pure in that no variable shares any common variance with any other in the explanatory set. The regression model with the orthogonal predictors is insensitive to the order in which the independent variables are entered into the equation. The explanatory power of each predictor is directly interpretable as the squared criterion correlation.

The transformation of the original variables is obtained from the matrix equation $Z^* = ZR - 1/2$ so that the correlation matrix (R^*) for the orthogonalized predictors in the columns of Z^* is $(Z^*)'Z^* / N = I$ the identity matrix (i.e., indicates orthogonality or 0 correlation), where Z is the 0 mean and standard

deviation of 1 (i.e., z score) matrix of original variables and $R^{-1/2}$ is the square root of the inverse of the correlation matrix R among the original variables. The correlation of the orthogonally transformed predictor variables with the original predictor variables is the square root of the correlation matrix R among the original predictor variables $R^{1/2}$.

The amount of predictable variance accounted for by each predictor variable in orthogonal regression is the correlation between the orthogonalized predictor and the criterion, and the overall fit of the model

(i.e., R^2) is simply the sum of the individual r^2 for each variable in the model.

—*John R. Reddon and James S. Ho*

See also Correlation Coefficient; Multicollinearity; Regression Analysis

Further Reading

Cawsey, T. F., Reed, P. L., & Reddon, J. R. (1982). Human needs and job satisfaction: A multidimensional approach. *Human Relations*, 35, 703–715.

P

I wanted to have a career in sports when I was young, but I had to give it up. I'm only six feet tall, so I couldn't play basketball. I'm only 190 pounds, so I couldn't play football. And I have 20–20 vision, so I couldn't be a referee.

—Jay Leno

PAGE'S L TEST

When comparing c populations in randomized complete block designs, researchers are often interested in examining ordered treatment effects among the c groups. The c levels of the treatment are either categorically ordered, as in the treatment condition “undergraduate class designation” with levels “freshman,” “sophomore,” “junior,” and “senior,” or the c levels are numerically ordered, as in the treatment condition “drug dosage” with levels “20mg,” “25mg,” “30mg,” and “35mg.” For situations where the researcher is interested in testing specifically for monotonically increasing or monotonically decreasing treatment effects, and either collects data only in ordinal form or collects data that are either interval- or ratio-scaled but does not desire to make the stringent assumptions necessary for employing a parametric procedure, the nonparametric L test devised by Ellis B. Page in 1963 may be selected.

Development

Let x_{ij} be the observed value under the j th treatment in the i th block (where $j = 1, 2, \dots, c$ and $i = 1, 2, \dots, n$).

In each of the n independent blocks, replace the c observations by their corresponding ranks such that a rank of 1 is given to the smallest observation in the block and a rank of c is given to the largest. Thus, R_{ij} is the rank (from 1 to c) associated with the j th treatment in the i th block.

The layout for the ranked responses from a sample of either n subjects or n blocks of matched subjects over c levels of a treatment condition are shown in Table 1.

Note that under the null hypothesis of no treatment effects, each ranking within a block is equally likely, so there are $c!$ possible rankings within a particular block and $(c!)^n$ possible arrangements of ranks over all n blocks. The sum of the ranks within each block is $c(c + 1)/2$, the sum of the ranks assigned to each of the c treatment levels. If the null hypothesis were perfectly true, the sum of the ranks for each of the c treatment levels would be $n(c + 1)/2$. On the other hand, if the alternative hypothesis is true and there is a perfect ordering in the c rankings among all n subjects, the sum of the ranks for each of the c treatment levels would respectively be $n, 2n, \dots, cn$ if the treatment effects were monotonically increasing, and $cn, (c - 1)n, \dots, n$ if the treatment effects were monotonically decreasing.

Table 1 Data Layout for the Page *L* Test

Block	Treatments				Totals
	1	2	...	<i>c</i>	
1	R_{11}	R_{12}	...	R_{1c}	$c(c+1)/2$
2	R_{21}	R_{22}	...	R_{2c}	$c(c+1)/2$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
<i>n</i>	R_{n1}	R_{n2}	...	R_{nc}	$c(c+1)/2$
Totals	$R_{.1}$	$R_{.2}$	...	$R_{.c}$	$nc(c+1)/2$

where

c = the number of treatment levels (i.e., columns)

n = the number of blocks (i.e., rows of subjects)

R_{ij} = the rank assigned to the j th treatment in the i th block

$R_{.j}$ = the sum of the ranks for treatment level j

To develop the Page test statistic L , the ranks assigned to each of the c treatment levels are totaled over all n blocks. That is, $R_{.j} = \sum_{i=1}^n R_{ij}$ is obtained for each of the treatments (where $j = 1, 2, \dots, c$). Page's test statistic L is

$$L = \sum_{j=1}^c j \cdot R_{.j}$$

where j is the hypothetical "ranking" (1 to smallest and c to largest) given to the j th treatment based on the alternative hypothesis of ordered treatment levels. Page shows that the test statistic L is approximately normally distributed with mean

$$\mu_L = \frac{nc(c+1)^2}{4}$$

$$\text{and standard deviation } \sigma_L = \sqrt{\frac{nc^2(c^2-1)(c+1)}{144}}.$$

For a nonparametric analysis, M_j represents some location parameter, typically the median of level j (where $j = 1, 2, \dots, c$). The Page *L* procedure may be used to test the null hypothesis of no treatment effect

$$H_0 : M_{.1} = M_{.2} = \dots = M_{.c}$$

against either the ordered alternative that a monotonically increasing treatment effect is present

$$H_1 : M_{.1} \leq M_{.2} \leq \dots \leq M_{.c} \\ \text{(with at least one inequality strict)}$$

or against the ordered alternative that a monotonically decreasing treatment effect is present

$$H_1 : M_{.1} \geq M_{.2} \geq \dots \geq M_{.c} \\ \text{(with at least one inequality strict)}$$

Note that the monotonic effects expressed by these two ordered alternative hypotheses are built into the calculation of the Page test statistic L , which correlates the hypothesized monotonic ordering of the c treatment levels with the sum of the ranks given by the n subjects in the c treatment levels. Thus, the Page *L* test statistic is related to Spearman's coefficient of rank correlation ρ .

To test the null hypothesis of no differences among the c treatment levels against the ordered alternative that differences among treatment levels are monotonically increasing, the decision rule is to reject the null hypothesis at an α level of significance if

$$Z \cong \frac{L - \mu_L}{\sigma_L} = \frac{L - \frac{nc(c+1)^2}{4}}{\sqrt{\frac{nc^2(c^2-1)(c+1)}{144}}} > Z_{1-\alpha},$$

where $Z_{1-\alpha}$ is the upper-tailed critical value from the standardized normal distribution. On the other hand, to test against monotonically decreasing treatment effects, the decision rule is to reject the null hypothesis at an α level of significance if $Z < Z_{\alpha}$ where Z_{α} is the lower-tailed critical value from the standardized normal distribution.

Applying Page's *L* Test

The following hypothetical example is based on a behavioral investigation of visual pattern recognition. The practical ramifications of the actual physiological research are in advertising and communications theory.

Suppose 10 subjects placed in front of a monitor are each presented with a single numerical digit (0–9) for a variable on-time of 10, 14, 18, or 22 milliseconds, immediately followed by the presentation of a randomly selected masking noise field that, for one-half second, may confound what digits the subjects had been shown. A computer program randomly generates the order in which the numerical digits appear as well as the particular on-time treatment level until 20 trials of each combination of digit and on-time level are observed by each subject. Based on this randomization process, each subject is given different sequences of digit and on-time level combinations until he or she completes the experiment.

To investigate the effects of different on-time levels on the percentage of correct responses, a randomized complete block design with but one observation per cell is used. The observation in each cell, x_{ij} , corresponds to the percentage of correct responses out of 200 trials under the j th on-time treatment level by the i th subject (where $j = 1, 2, \dots, c = 4$ and $i = 1, 2, \dots, n = 10$).

Given these different numerical on-time levels, it may be of primary interest for the researcher to test for ordered treatment effects. Transforming each subject's percentage of correct responses across the four treatment levels into ranks, the Page *L* test can be employed.

The original data values x_{ij} along with the transformed ranks R_{ij} are displayed in Table 2.

To test the null hypothesis of no treatment effect (that is, each of the four on-time levels results in similar pattern recognition abilities across subjects, and any observed differences are due to chance),

$$H_0 : M_{.1} = M_{.2} = M_{.3} = M_{.4}$$

against the ordered alternative that a monotonically increasing treatment effect is present

$$H_1 : M_{.1} \leq M_{.2} \leq M_{.3} \leq M_{.4} \\ \text{(with at least one inequality strict)}$$

Table 2 Hypothetical Results of Visual Pattern Recognition Experiment

Subjects	On-Times in msec				Totals
	10	14	18	22	
1=AA	63% → 1	69% → 2	76% → 4	75% → 3	10
2=AB	38% → 2	37% → 1	40% → 3	50% → 4	10
3=KB	43% → 1	45% → 2	47% → 3	53% → 4	10
4=LB	54% → 4	51% → 2	48% → 1	52% → 3	10
5=ED	43% → 2	48% → 4	41% → 1	47% → 3	10
6=KG	29% → 3	26% → 1	28% → 2	36% → 4	10
7=RK	69% → 1	75% → 2	79% → 3	85% → 4	10
8=DM	48% → 2	45% → 1	61% → 4	55% → 3	10
9=MM	17% → 1	20% → 2	22% → 3	30% → 4	10
10=GT	59% → 3	63% → 4	56% → 1	57% → 2	10
Rank Totals	20	21	25	34	100

Note: Each cell contains the measured response and the transformation to ranks.

the Page *L* procedure may be used. From the rank sums over the j treatment levels in Table 2, the Page *L* statistic is computed as follows:

$$L = \sum_{j=1}^c j \cdot R_{.j} \\ = (1)(20) + (2)(21) + (3)(25) + (4)(34) \\ = 273.$$

Moreover,

$$\mu_L = \frac{nc(c+1)^2}{4} = \frac{(10)(4)(5^2)}{4} = 250$$

and

$$\sigma_L = \sqrt{\frac{nc^2(c^2-1)(c+1)}{144}} \\ = \sqrt{\frac{(10)(4^2)(4^2-1)(5)}{144}} \\ = 9.129.$$

To perform the Page *L* test,

$$Z \cong \frac{L - \mu_L}{\sigma_L} = \frac{273 - 250}{9.129} = 2.52.$$

Using a .05 level of significance, given that $Z = 2.52 > Z_{.95} = 1.645$, the upper-tail critical value from the standardized normal distribution, the null hypothesis is rejected. A significant ordered treatment effect is present. The p value for Page's *L* test is 0.0059.

Discussion

In order to examine a set of c ordinal responses pertaining to the levels of some treatment condition provided either by a sample of subjects used as their own controls to make these assessments or by a sample of c matched subjects randomly assigned to each treatment level as members of a block, in 1937 the economist Milton Friedman developed a non-parametric competitor of the two-way randomized block *F* test. This test should be useful for testing against general alternatives when the data collected are only ordinal or when the researcher does not desire to make the more stringent assumptions underlying the classical parametric ANOVA procedure.

When employing the Friedman test, a rejection of the null hypothesis of no differences in the c treatment levels leads to the generalized conclusion that a treatment effect is present. This implies that at least one of the groups is different from another group or from all other groups. Owing to the general nature of the conclusions that would be drawn, it becomes imperative for the researcher to employ a posteriori an appropriate multiple comparisons procedure in order to determine where the actual significant differences in the treatment levels are.

The Page *L* test is, in a sense, a nonparametric extension of the more general Friedman test. If interest centers on global or generalized findings, the Friedman should be selected. However, if the treatment levels are ordered and interest is in monotonic effects, the Page *L* test is much preferred.

Comment: The Impact of Testing for Ordered Alternatives

It is essential to a good data analysis that the appropriate statistical procedure be applied to a specific situation. If the Friedman test were used on the data

displayed in Table 2, the null hypothesis of no differences in treatment effects

$$H_0 : M_{.1} = M_{.2} = \dots = M_{.c}$$

would be tested against the more general alternative

$$H_1 : \text{Not all } M_j \text{ are equal (where } j = 1, 2, \dots, c).$$

From Table 2, it is observed that $n = 10$, $c = 4$, and the rank sums for the four treatment levels are 20, 21, 25, and 34, respectively.

For these data, the Friedman test statistic $T = 7.32 < \chi^2_{0.05, (c-1=3)} = 7.815$, the upper-tailed critical value under the chi-square distribution with 3 degrees of freedom. Thus, it is important to note that in using a .05 level of significance, the null hypothesis is not rejected. The Friedman test fails to detect a significant treatment effect, and no a posteriori analysis on the $c = 4$ levels can be undertaken. The p value for the Friedman test is 0.0624.

Conclusions

The Page *L* test is quick and easy to perform. The only assumptions are that either the n subjects providing the repeated measurements (i.e., the ranks) across the c treatment levels are independently and randomly selected, or the n blocks of homogeneous subjects examining the c treatment levels are independently and randomly selected.

When evaluating the worth of a statistical procedure, statistician John Tukey defined "practical power" as the product of statistical power and the utility of the statistical technique. Based on this, the Page *L* test enjoys a very high level of practical power under many useful circumstances. Although some competing tests may provide slightly more statistical power than the Page *L* test, no other test of this type can match its distinct simplicity and computational ease.

—Mark L. Berenson

Further Reading

Berenson, M. L. (1982). Some useful nonparametric tests for ordered alternatives in randomized block experiments. *Communications in Statistics: Theory and Methods*, 11, 1681–1693.

- Berenson, M. L. (1982). A study of several useful nonparametric tests for ordered alternatives in randomized block experiments. *Communications in Statistics: Simulation and Computation*, 11, 563–581.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675–701.
- Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216–230.
- Tukey, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics*, 1, 31–48.

PAIRED SAMPLES *t* TEST (DEPENDENT SAMPLES *t* TEST)

Paired samples *t* test, also known as dependent samples *t* test, is used when there are two groups to compare, wherein the scores in one group are linked or paired with scores in the other group. In this situation, the assumption of independence has been violated, so an independent *t* test cannot be used.

An example of when to use a paired samples *t* test is a study examining happiness of twins (based on a survey with a composite score ranging from 1 to 100), with both twins involved in the study. Because the scores are paired or dependent (each twin's score is related to the other twin's score), these data would be analyzed using a paired samples *t* test. This is the best choice of statistic to use because the scores from one twin are linked, or might be similar, to the scores from the other twin.

Another example of when it would be appropriate to use the paired samples *t* test is with repeated measures. For example, to understand if a teaching method is effective, a pretest could be given on the first day of class and then again on the last day of class. The scores from these tests (the data) would be considered linked for each student. Therefore, in the case of repeated measures with two sets of scores, the paired samples *t* test would be an appropriate choice for a statistic.

Table 1 shows an example of data that would be appropriate on which to use the paired samples *t* test. The first variable is “twin 1 happiness scores” and the

Table 1 Data of Twin 1 and Twin 2 Happiness Scores

Twin 1 Happiness Scores	Twin 2 Happiness Scores
88	92
75	84
45	52
95	90
50	52
79	80
69	75
48	50
59	64
58	58

second variable is “twin 2 happiness scores.” Happiness is measured on a scale of 1 to 100, where 1 is very unhappy and 100 is extremely happy.

Assumption of the Paired Samples *t* Test

There is one important assumption or condition for the paired samples *t* test: The variables should be normally distributed. This can be tested with a computer program such as SPSS with the skewness and kurtosis values with the Explore command. A graphical representation of the normal distribution can be obtained through the Q-Q plot.

If the assumption of normality is not met, the Signed Rank Test should be computed instead.

Research Hypothesis

The null hypothesis analyzed with the paired samples *t* test is similar to the hypothesis used with the independent samples *t* test:

$$H_0 : \mu_1 = \mu_2$$

This hypothesis is testing that the means are equal. The alternative hypothesis would be that the means are not equal:

$$H_a : \mu_1 \neq \mu_2.$$

Computing the Value for the Paired Samples *t* Test

The formula for the paired samples *t* test is as follows:

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}},$$

where $\bar{X}$ is the mean for the first variable and $\bar{Y}$ is the mean for the second variable. The denominator, $s_{\bar{D}}$ is the standard error of the difference between the means. It is calculated with the following formula:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}}, \text{ where } s_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N - 1}}.$$

The $s_{\bar{D}}$ is the standard error of the difference between the means. The s_D is the standard deviation of the difference between the means, the D is the difference between each paired score ($X - Y$), and N is the number of paired scores.

Using the Computer to Compute the Value

The data used in the above example were entered into SPSS. To check whether the data for each variable are normally distributed, the skewness values can be computed through the Explore command and the Q-Q plot can be generated in the computer program SPSS. The output is in Figures 1 and 2.

Next, the paired samples *t* test is computed with the Paired Samples *t* Test command found under Compare Means. The output is in Figure 3. As shown in the last table of the output, there is a statistically significant difference between the twins' happiness scores, $t(9) = -2.44$, $p = .037$, $d = .77$. By examining the means in the first table, we see that Twin 2 has a higher

mean happiness score ($M = 69.70$) than Twin 1 ($M = 66.60$). The effect size is calculated by $d = \frac{\text{Mean}}{SD}$, where the *Mean* is the paired differences mean and *SD* is the paired differences standard deviation. Thus, for this example, $d = \frac{3.1}{4.012} = .77$.

There are advantages to using dependent samples. Dependent samples tend to reduce the effects of variability between the elements in a sample that can confound the results. For example, suppose the sample obtained is an independent sample of students, one group of students who took a course and one group who did not. One of the students who did not take the course has an extremely high score. If we examine the sample means of the scores for each group, this student's score might distort the results, thus making the sample mean for the students who have not taken the course unrealistically high. Dependent samples

Descriptives				
			Statistic	Std. Error
twin 1 happiness scores	Mean		66.60	5.492
	95% Confidence Interval for Mean	Lower Bound	54.18	
		Upper Bound	79.02	
	5% Trimmed Mean		66.22	
	Median		64.00	
	Variance		301.600	
	Std. Deviation		17.367	
	Minimum		45	
	Maximum		95	
	Range		50	
	Interquartile Range		32	
	Skewness		.346	.687
	Kurtosis		-1.194	1.334
twin 2 happiness scores	Mean		69.70	5.198
	95% Confidence Interval for Mean	Lower Bound	57.94	
		Upper Bound	81.46	
	5% Trimmed Mean		69.56	
	Median		69.50	
	Variance		270.233	
	Std. Deviation		16.439	
	Minimum		50	
	Maximum		92	
	Range		42	
	Interquartile Range		34	
	Skewness		.097	.687
	Kurtosis		-1.836	1.334

Figure 1 Output From the Explore Command in SPSS

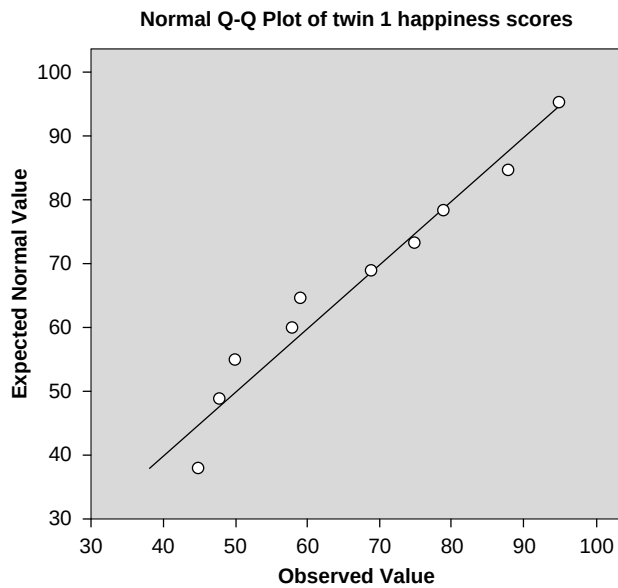


Figure 2 Q-Q Plots Generated With the Computer Program SPSS

would create a different situation. In this example, the data would include a “before the course” and “after the course” score for each student. If one student in the study scores high, this should appear in both samples and therefore will not distort just one of the samples. In other words, things would balance out. Thus, the effects of variability among the participants in the sample will be reduced.

Whereas some studies involve using the same group of individuals with measurements recorded before and after some intervening treatment, some other studies use naturally occurring pairs by matching on factors with an effect that might otherwise obscure differences (or the lack of them) between the two populations of interest. Paired samples often provide more information than would independent samples because extraneous effects are ruled out. It is important to remember that the matching needs to be

Paired Samples Statistics				
	<i>Mean</i>	<i>N</i>	<i>Std. Deviation</i>	<i>Std. Error Mean</i>
Pair 1 twin 1 happiness scores	66.60	10	17.367	5.492
1 twin 1 happiness scores	69.70	10	16.439	5.198

Paired Samples Correlations				
		<i>N</i>	<i>Correlation</i>	<i>Sig.</i>
Pair 1	twin 1 happiness scores & twin 2 happiness scores	10	.973	.000

Paired Samples Test									
		Mean	Paired Differences				t	df	Sig.(2-tailed)
			Std. Deviation	Std. Error Mean	95% confidence Interval of the Difference				
					Lower	Upper			
Pair 1	twin 1 happiness scores twin 2 happiness scores	−3.100	4.012	1.269	−5.970	−.230	−2.443	9	.037

Figure 3 Paired Samples *t* Test Output Generated With the Computer Program SPSS

undertaken accurately, because the samples will then be considered equal.

—Nancy L. Leech, Anthony J. Onwuegbuzie,
and Larry Daniel

Further Reading

- Glass, G. V. & Hopkins, K. D. (1995). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson Education.
- Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). *SPSS for basic statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
- Urdan, T. C. (2001). *Statistics in plain English*. Mahwah, NJ: Erlbaum.

SPSS tutorial: How to do a paired samples *t* test: <http://academic.uofs.edu/departments/psych/methods/cannon99/level2c.html>

PAIRWISE COMPARISONS

For psychological variables, the operations of addition and multiplication—and therefore most elementary statistics—cannot be applied to scales that are constructed from pairwise comparison data. Pairwise comparisons may be used to construct ordinal scales, but because the operations of addition and multiplication are not applicable to ordinal data, the construction of scales from ordinal data is problematic, especially in the multicriteria and multiple decision makers cases. In those cases where the construction of ordinal scales is possible, the operations of addition and multiplication are not applicable to scale values.

Operations on Scale Values

The application of elementary statistics—such as standard deviation—to scale values requires the availability of the operations of addition and multiplication as well as order and the limit operation of calculus. Psychological variables to which addition, multiplication, order, and the limit operation are applicable must be modeled in the same manner—and for the same mathematical reasons—as such familiar physical

variables as *time*, *position* of points on a straight line, *potential energy*, and *temperature* on the Fahrenheit or Celsius (but not Kelvin) scales.

Reference Objects

The building blocks for such scales require three or four rather than two objects. Consider, for example, the statement that the temperature of a certain object is 100 degrees on the Fahrenheit scale. As can be seen in Figure 1, this statement involves three empirical objects, three mathematical objects, and three correspondences: The empirical objects are freezing water, boiling water, and the object under measurement; the mathematical objects are the numbers 32, 100, and 212; and the correspondences are the assignments of the temperatures—{freezing water, 32}, {object under measurement, 100}, and {boiling water, 212}. It should be noted that this statement requires two empirical reference objects (freezing water and boiling water) and two corresponding mathematical reference objects (the numbers 32 and 212).

Removing one of these empirical reference objects and its corresponding mathematical reference object results in an ordinal pairwise comparison where neither differences nor ratios are defined. If both empirical reference objects are removed and the numerical ones are not, the statement “on a scale of 32 to 212, an object scores 100” is obtained. Statements of this form, such as the common phrase “on a scale of 1 to 10, an object scores 7,” have no mathematical meaning. No mathematical operations or statistics are applicable to numbers produced from such statements.

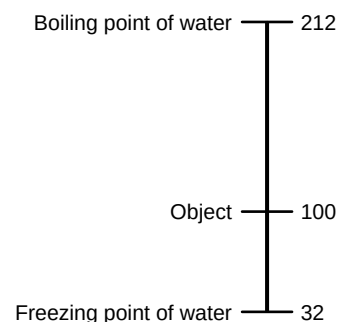


Figure 1 Example of Scale Values and Reference Objects

Ratios

Ratios of the type T_1/T_2 have become defined for *temperature* only after it has been established that temperature has an absolute zero. Conversely, for variables where the existence of an absolute zero has not been established, such ratios are undefined. For example, for *time*, the ratio t_1/t_2 where t_1 and t_2 are two points in time is undefined, whereas the ratio of two *time differences*, (i.e., time periods or time intervals) $(\Delta t)_1/(\Delta t)_2$, is well-defined. It follows that the ratio v_1/v_2 is undefined for any psychological variable because the existence of an absolute zero has not been established for such variables. In particular, decision methodologies such as the Analytic Hierarchy Process that depend on data in the form of ratios of preferences are not valid methodologies. In general, a ratio of differences depends on four points, but this ratio may depend on three variables rather than four when two of the four points are identical. Although the number of variables in this expression can be reduced from four to three, it cannot be further reduced to a pairwise comparison.

Ordinal Comparisons

Pairwise comparisons are applicable to ordinal data, but the operations of addition and multiplication and the concepts of cancellation and trade-off do not apply to ordinal scales. For example, ordinal methodologies ignore the question “By how much is Object A preferred to B?” because the concepts of *difference*, *slight difference*, *large difference*, or “*twice as preferable*” are inapplicable in the ordinal case.

Mathematical Structure

In technical terms, in the case of physical or mathematical variables, for addition, multiplication, order, and limits to be applicable, the measured objects must correspond to scalars in the ordered complete field of real numbers, to vectors in a one-dimensional vector space over this field, or to points in a one-dimensional affine space over the real numbers. The zero vector in a vector space is an absolute zero because it is a fixed point of the automorphisms of the space. For psychological variables where the existence of an absolute zero is not established, the only possibility for addition,

multiplication, order, and limits to be applicable is the model where the measured objects correspond to points in a one-dimensional affine space over the ordered real numbers. In such a space, the ratio of two points is undefined, whereas their difference is a vector and the ratio of two vectors is a scalar.

—Jonathan Barzilai

Further Reading

- Barzilai, J. (1998). On the decomposition of value functions. *Operations Research Letters*, 22, 159–170.
- Barzilai, J. (2001). Notes on the analytic hierarchy process. *Proceedings of the NSF Design and Manufacturing Research Conference* (pp. 1–6). Tampa, Florida.
- Barzilai, J. (2004). Notes on utility theory. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1000–1005.
- Barzilai, J. (2005). Measurement and preference function modelling. *International Transactions in Operational Research*, 12, 173–183.
- Barzilai, J. (2006). Preference modelling in engineering design. In K. Lewis, W. Chen, & L. Schmidt (Eds.), *Decision making in engineering design*. New York: ASME Press.

Jonathan Barzilai Web page: <http://myweb.dal.ca/barzilai/>

PARALLEL COORDINATE PLOTS

Parallel coordinate plots were first introduced by Inselberg and by Wegman. The main idea of parallel coordinate plots is to switch from Cartesian coordinates, where points are plotted along orthogonal axes,

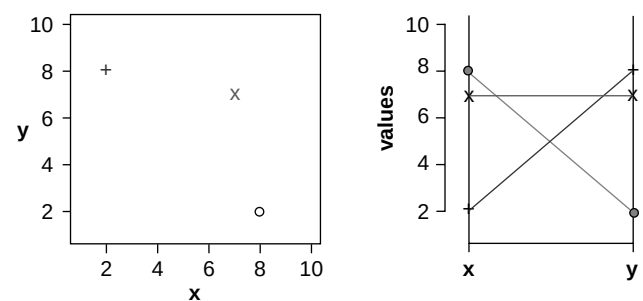


Figure 1 Relationship Between Scatterplot and Parallel Coordinate Plots in a 2D Data Sketch

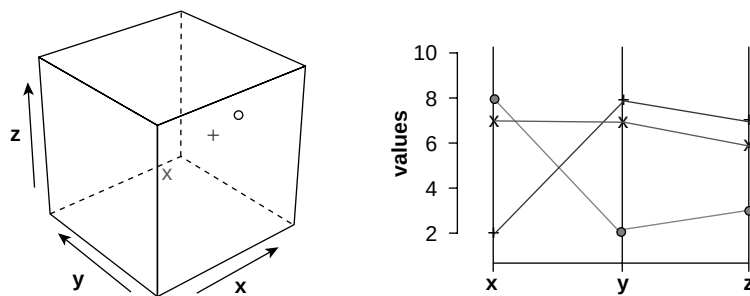


Figure 2 Relationship Between a 3D Scatterplot and Parallel Coordinate Plots in 3D Data

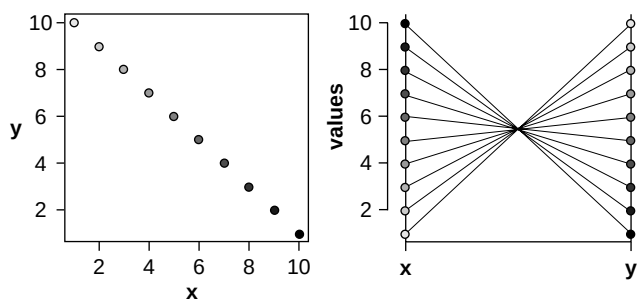


Figure 3 Scatterplot and Parallel Coordinate Plot of 10 Points on a Line (Left)

Note: Left line translates to the common intersection point on the right.

to a projective geometry, where axes are plotted side by side.

Figures 1 and 2 show this principle for three data points in two and three dimensions. On the left-hand side of Figure 1, the scatterplot shows the three points (2, 8), (7, 7), and (8, 2) plotted as +, x, and o, respectively. The same data points are plotted on the right-hand side: The first observation (2,8) is shown by first putting a “+” at the 2 of the x-axis, then a “+” at the value 8 of the y-axis. Both of these markers are then connected by a line. This way, each observation corresponds to a single line.

The advantage of this approach becomes visible in Figure 2. Measurements in a third variable are added to our previous example. Although we would need interactive tools, such as rotation, to be able to see the

relationship between the three points in the three-dimensional scatterplot on the left, the parallel coordinate plot on the right is extended naturally to three dimensions by putting another axis alongside the other two, drawing markers for each of the measurements and connecting them by lines to the corresponding values of the neighboring axis.

Obviously, this principle is applicable far beyond the (usually up to three) dimensions of Cartesian displays.

The disadvantage of parallel coordinate plots is that we lose our familiar coordinate system. Because the two systems show the same information, it is a matter of a little practice to get used to the new coordinate system.

Properties

The basic properties of a principal coordinate plot are based on the duality between projective and Euclidean geometry: Points in one system correspond to lines in the other. The scatterplot on the left-hand side of Figure 3 shows 10 data points on a straight line. The same points are shown in the parallel coordinate plot on the right. All 10 lines meet in the same point, indicating the linear relationship.

Example

Parallel coordinate plots are used mostly to find and explore high-dimensional clusters in data. Figure 4

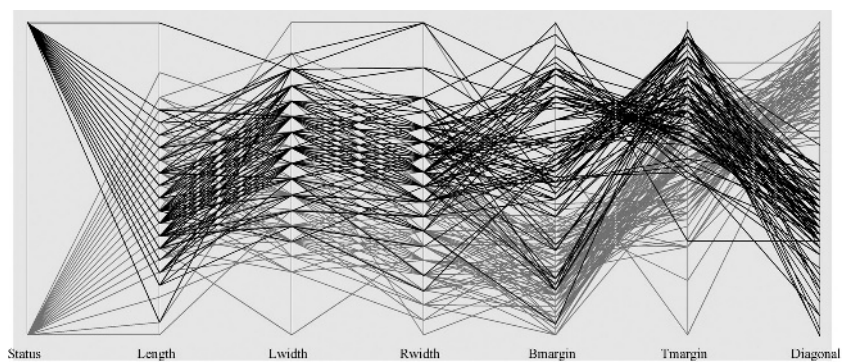


Figure 4 Parallel Coordinate Plot of the Swiss Bank Note Data

Note: Forged bank notes are marked in black.

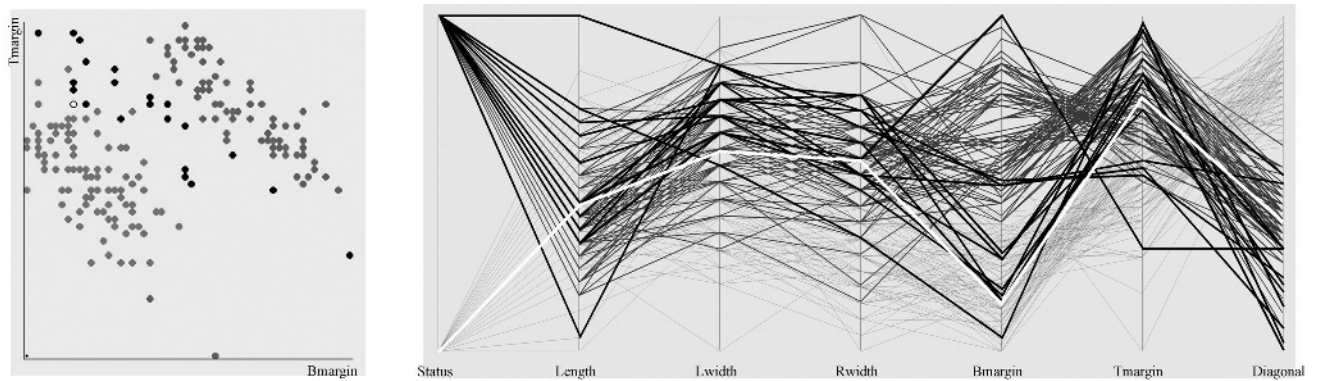


Figure 5 Two Different Classes of Forgeries

Note: The two classes of forgeries (marked in dark grey and black, respectively) are visible in the Swiss bank note data set. Marked in white is another bill. From its measurements it fits very well into the smaller group of forged bills. It was classified as genuine by the bank, however.

shows a parallel coordinate plot of the Swiss bank note data set by Flury and Riedwyl. Measurements of 100 genuine and 100 forged Swiss bank notes were taken, with the goal of identifying forgeries. Six different measurements are available to us: the horizontal length of a bill at the top (Length), left and right vertical width (Lwidth, Rwidth), vertical width of the margin at center bottom (Bmargin) and center top (Tmargin), as well as the diagonal length of the image on the bill. Marked in grey are genuine bills. From the parallel coordinate plot, we can see that the forged bills tended to have larger margins at the bottom and top, and the diagonal length of the image seemed to be smaller than those measurements of genuine bills.

Among the forgeries, we can find a “knot” between the axes BMargin and Tmargin, indicating a negative linear relationship between bottom and top margin. What is strange, though, is a very tiny second knot among the forgeries between the same axes a little bit further down closer to the genuine bills. The smaller knot is marked in black in Figure 5. The two knots clearly identify two different types of forged bills. Marked by the unfilled circle on the left and the white line on the right of Figure 5 is another bill, classified as genuine; from its overall measurements, it seems to fit into the second class of forgeries very well.

—Heike Hofmann

Further Reading

- Asimov, D. (1985). The grand tour: A tool for viewing multi-dimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6, 128–143.
- Carr, D., Wegman, E., & Luo, Q. (1997). *ExplorN: Design considerations past and present*. Technical Report 137, Center for Computational Statistics, George Mason University, Fairfax, VA.
- Cook, D. (1997). Calibrate your eyes to recognize high-dimensional shapes from their low-dimensional projections [Electronic version]. *Journal of Statistical Software*, 2(6).
- Flury, B., & Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. New York: Chapman & Hall.
- Inselberg, A. (1985). The plane with parallel coordinates. *Visual Computer*, 1, 69–91.
- Inselberg, A. (1999). Don't panic . . . just do it in parallel! *Computational Statistics*, 14, 53–77.
- Wegman, E. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85, 664–675.
- Wegman, E. J., & Luo, Q. (1997). High dimensional clustering using parallel coordinates and the grand tour. *Computing Science and Statistics*, 28, 361–368.

The Apoala project is an offspring of GeoVista and is a dynamic parallel coordinate plot, implemented in TCL, designed to show the relationships between multiple variables in large data sets: <http://www.geovista.psu.edu/products/demos/dsall/Tclets072799/pcpdescription.html>
Parallel coordinate plots software, including CASSATT, a stand-alone Java application: <http://www.rosuda.org/software/>

ExplorN software for touring through high-dimensional data using parallel coordinate plots: <ftp://www.galaxy.gmu.edu/pub/software/> (It is now replaced by its commercial evolution, CrystalVision)

PARALLEL FORMS RELIABILITY

All observations and all measurements contain error. The focus of much work in measurement is to minimize and estimate the amount of error in any given measurement. In classical test theory, X is an observed score that is composed of T , the true score, and E , the error score: $X = T + E$. The true score is never known, but can be thought of as the long-range average of scores from a single instrument administered to an individual an infinite number of times (the expected value or expected score). The error score is random and may have many sources, including testing conditions, individual characteristics that fluctuate from administration to administration, differences in forms, or instability of an individual's ability or trait over time.

This random error score is quite different from systematic sources of error, like testwiseness, which may systematically increase an individual's score on each administration. Because testwiseness is systematic or constant, it finds its way into the true score and creates problems regarding validity, in that the trait being measured inadvertently may be influenced by testwiseness. Random error, because it varies randomly, influences the consistency of scores, but not the expected value of a score (the true score), and thus influences reliability, not validity.

Theoretically, we can estimate the amount of error if we know how much of a given score is due to errors of measurement. If we were able to test a single person repeatedly without the effects of recall and fatigue, variation in his or her scores would be considered measurement error. If there was no measurement error, he or she would get the same score on each administration. Because it is not possible to test individuals repeatedly without the interference of recall and fatigue, we employ groups to estimate measurement error variance. This allows us to estimate the standard error of measurement, the typical amount of measurement error in a set of scores.

If we take the classical test theory model of scores and consider groups of scores and their variances, we see that the variance of the observed scores equals the sum of the variance of true scores and the variance of error scores: $S_X^2 = S_T^2 + S_E^2$ (in sample notation).

This is the long way of introducing the need for reliability; reliability is a tool used to estimate the standard error of measurement, but it also has some intrinsic benefits in and of itself. Theoretically, reliability is considered the correlation between scores on two parallel forms of a test. The idea is that if there is no measurement error at work, scores from two parallel forms administered to the same group of individuals should be perfectly correlated—each individual should obtain the same score. It can be shown that the correlation between two parallel forms of a test is equal to the ratio of true score variance to observed score variance—the proportion of variance in observed scores that is due to true individual differences: $r_{tt} = \frac{S_T^2}{S_X^2}$. This reliability coefficient then can be used in estimation of the standard error of measurement, because it tells us the proportion of observed variance that is true variance; the standard error of measurement is a function of the proportion of observed variance that is true variance.

Estimating Reliability

Based on the classical test theory conception of reliability, scores are reliable to the extent that individuals' scores remain constant on repeated measurements. One way to estimate the degree of consistency of scores is to employ parallel forms—two or more forms of a measurement instrument that are built to the same specifications in terms of the content domain or construct definition. Each form provides a separate sample from the domain.

This form of reliability is best estimated when the actual measurement procedure employs multiple forms, because if scores vary between forms for a single individual, we are likely to attribute this to differences in content—error results from item or content sampling error. This is typically done by administering two forms to each individual during the same administration time. To avoid order effects, subjects can be randomly assigned one form first and the other

second. The scores resulting from the two forms are then correlated; this correlation is the parallel forms reliability coefficient.

Formal Conditions for Parallel Forms Reliability

Parallel forms reliability is estimated from the administration of two forms simultaneously. This is appropriate when the measurement procedure includes multiple forms or when the inference from scores is based on the ability to generalize in terms of the items selected for a given form. Knowledge of parallel form reliability allows one to generalize from one sample of items to a larger domain of items. To be a coefficient of reliability, the parallel forms should be built to the same content specifications and have equal means, variances, and item intercorrelations.

This form of reliability is also called alternate forms reliability or the coefficient of equivalence. This form of reliability can be estimated from forms that are parallel or equivalent, where equivalent forms are similar in content but not statistical quality. Note that if scores are uniformly different between two forms by a constant (perhaps because one test is more difficult than the other), the correlation between two scores is potentially perfect (1.0). This does not mean that the measures are exchangeable in any practical sense; such tests are often equated to adjust for differences in difficulty. In this context, the parallel conditions of the forms have not been met (form means are not equal), and the reliability coefficient or resulting correlation is more difficult to interpret.

General Issues Regarding Reliability

Because the parallel forms estimate of reliability is based on a correlation, it is not simply a characteristic of the measurement instrument itself. Score variability directly influences correlations, such that all else being equal, the more score variance present, the higher the correlation and thus the higher the reliability. Correlational forms of reliability are sample specific and thus not necessarily generalizable to other samples. They do, however, provide an estimate of score consistency for the scores at hand.

In any estimate of reliability, conditions present during the specific administration of the measurement instrument can affect performance and scores in random ways, leading to lower consistency of scores and lower reliability. Each type of reliability estimate (e.g., parallel forms reliability) also captures a specific form of random error. The parallel forms reliability primarily captures measurement error due to sampling items from a domain, a form of sampling error. If this source of error is important to estimate given the measurement procedure (because multiple forms are used and we assume they are equivalent random samples of the domain), then it is an appropriate form of reliability. Technically speaking, an estimate of reliability should be obtained for each set of scores because any one estimate is sample specific, and the argument of generalizability across samples is difficult to make.

Finally, because sampling error is a function of sample size, all else being equal, longer forms will yield higher reliability coefficients. Better, larger samples of items from the domain will reduce the likelihood that two forms differ in their ability to cover the domain. There is a functional relation between form length and reliability that is represented by the Spearman-Brown prophecy formula.

—Michael C. Rodriguez

See also Coefficient Alpha; Reliability Theory; Standard Error of Measurement

Further Reading

- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education; Macmillan.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

PARAMETER

A parameter is a characteristic that describes the elements in a population. It represents a collective

measure that summarizes the data elements, which comprise the population. Such measures may indicate the central tendency, dispersion or variability, symmetry, or the degree of peakedness in the distribution of the data points in the population. Because a parameter is always associated with a population, it is vital to define the population first.

Population

A population is the set of all items possessing a characteristic of interest, as defined by the problem. Suppose the objective is to study the distribution of income of all working people in a given state. The population in this case is the set of all working people (with an income) in the chosen state.

Parameter Examples

In relation to the population of all working people in the selected state, a parameter that represents the central tendency of the income distribution could be the mean income or the median income. Typically, parameters are denoted by Greek letters. For example, the mean income of all working people in the chosen state could be \$40,000, which would be represented by μ .

Another example of a parameter that represents the variability in the income could be the range. The population range is defined as the difference between the maximum and minimum value, when all elements in the population are considered. Hence, the population range could be \$500,000, a measure that depicts the spread in the population values.

Note that in order to come up with the value of the parameter (i.e., population mean μ), one needs to take into consideration each and every data point in the population. Sometimes, this may be a time-consuming and costly process, as it may be accomplished only through a census. The population mean is calculated as follows:

$$\mu = \frac{\sum_{i=1}^N X_i}{N},$$

where

X_i denotes the value of the i th element in the population;

Σ , the Greek letter sigma, is the summation sign, which directs us to add the value of all the elements (1, 2, . . . , N) in the population;

N represents the total number of elements in the population, and is also known as the population size.

Differences Between Parameter and Statistic

Whereas a parameter represents a characteristic of a population, a statistic represents a characteristic of a sample chosen from the defined population. The value of a parameter remains constant, for a static problem, whereas the value of a statistic may change from sample to sample.

Consider the previously defined problem of income distribution of working people in a state for the month of September 2005. Suppose there were 5,000 working people during that month. The population mean, μ , would be the average income of all 5,000 people for that month, which would be a constant value. On the other hand, if a random sample of 500 people is selected from the population, and the average income of those in the sample is calculated, it would be a statistic. The sample mean, $\bar{X}$, a statistic, could be \$38,000, which necessarily may not equal the value of the corresponding parameter, μ . Furthermore, the value of the sample mean will usually change from sample to sample.

In practice, because a census is not always feasible, a sample statistic is used to estimate the value of a corresponding population parameter. Thus, the sample mean, $\bar{X}$, serves as an estimator of the population mean, μ .

—Amitava Mitra

Further Reading

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2006). *Introduction to probability and statistics* (12th ed.). Pacific Grove, CA: Brooks/Cole.

Parameter definition: <http://www.animatedsoftware.com/statglos/sgparam.htm>

Parameters: <http://davidmlane.com/hyperstat/A12328.html>

PARAMETER INVARIANCE

According to *Webster's New Universal Unabridged Dictionary*, *invariant* simply means “constant,” and, in mathematics more specifically, “a quantity or expression that is constant throughout a certain range of conditions” (p. 1003). Furthermore, a *parameter* in a statistical model is an explicit component of a mathematical model that has a value in one or multiple populations of interest; its value is estimated with a certain estimation routine based on sample data to calibrate the model. Therefore, *parameter invariance* simply means that a certain parameter in a certain statistical model is constant across different measurement conditions such as examinee subgroups, time points, or contexts.

From a decision-making viewpoint, parameter invariance is a desirable characteristic of a statistical model, because it implies that identical statistical decisions can be made across different measurement conditions. In this sense, parameter invariance is one of the preconditions for the generalizability of inferences.

Four aspects of parameter invariance warrant a separate highlighting because they are easily overlooked by practitioners. First, parameter invariance is an abstract ideal state, because parameters are either invariant or not. If a parameter varies across measurement conditions, then a *lack of invariance* is present, which is a continuum of differing values. However, it is always open to debate and secondary assessment whether the difference between parameter values across conditions affects practical decision making. Second, parameter invariance is always tied to a specific statistical model that contains parameters as elements. Therefore, the issue of whether parameters are invariant cannot be answered abstractly without reference to a certain statistical model. Third, parameter invariance implies a comparison of parameter values

across measurement conditions, making discussions about invariance without a reference frame for what constitutes different conditions similarly irrelevant.

The Mathematical Formalization of Parameter Invariance

In the psychometric literature, particularly in the area of confirmatory factor analysis, parameter invariance has had a long tradition, and important groundwork was laid in the middle of the previous century. In the early 1990s, when models in Item Response Theory (IRT) became more popular and accessible, the work of Meredith and colleagues expanded the rich basis of work by Jöreskog and McDonald. At the end of the 1990s, procedures for assessing parameter invariance had become fully integrated into many software programs for latent variable models, such as BILOG-MG, Mplus, LISREL, or AMOS.

Today, parameter invariance has also become fully conceptually integrated into comprehensive and unifying treatments of modeling frameworks in the area of psychometrics and, sometimes, does not appear to have any particular importance for authors that merits a treatment in separate chapters or even books.

Methodological Approaches for Assessing Parameter Invariance

The following common approaches are organized by the type of comparison that is being done (i.e., across subgroups of items, persons, or time points). Moreover, this entry will focus on parameter invariance for psychometric assessment as it is investigated with both observed- and latent-variable models. Although the issue of what constitutes a latent variable is complex itself, for the purpose of this entry, latent variable models are defined as statistical models that contain continuous or discrete unobservable variables whose values are estimated with a specific algorithm.

Parameter Invariance Across Items

In classical test theory (CTT), the total test score is used to rank-order examinees or to make mastery

decisions about them, and its value typically gets supplemented with information about its precision at different score levels. Moreover, measurement models can be specified in terms of how the true scores of different items are related to each other and how error variances are related to one another (e.g., parallel items, τ -equivalent items, congeneric items). Within the framework of structural equation modeling (SEM), the assumptions that these different measurement models postulate about the true score relationship, error variances, and covariances can be estimated today. The added analytical power, however, comes at the price of requiring data structures that make the estimation of such models computationally feasible.

Parameter Invariance Across Examinee Subgroups

The most common setup for investigating parameter invariance is one where the properties of an assessment instrument are compared for different subgroups of examinees such as different ethnic or gender groups. It is important to highlight that the definition of these groups is by no means “natural” in any common sense of the term; rather, it is typically driven by either legislative or theoretical needs. Alternatively, one can view the problem of which groups to compare as an empirical one such that the grouping structure should be data driven. In this case, one works within the area of latent class models, where models such as a mixture Rasch model are employed. Of course, additional parameters such as group membership proportions get estimated along the way, but if the same measurement model is postulated across groups, the core of the model is concerned with parameter invariance.

If parameters for individual items are investigated for invariance, such research is more commonly known under the label *differential item functioning* (DIF); for sets of items (i.e., item bundles or testlets), such research is known under the labels *differential bundle functioning* (DBF) and *differential test functioning* (DTF), respectively. A wide range of techniques for assessing DIF have been proposed, some of

which condition on manifest variables such as the observed total score, whereas others condition on latent variables such as θ in an IRT model.

The first class of models includes procedures such as the traditional Mantel-Haenszel χ^2 statistic, loglinear models, logistic regression models, and discriminant function analysis models. The second class of models leads to a variety of approaches that compare, for example, the area between item response curves for dichotomous and polytomous scoring models or compute an adjusted weighted difference in proportions of examinees at different score levels, as in the SIBTEST, Poly-SIBTEST, and MULTISIB packages.

Common to all of the procedures is that a variety of decisions have to be made to implement them in practice. Specifically, examinees need to be matched on a criterion so that the performance of the different groups of interest can be conducted for people with identical characteristics on the construct that is to be measured. This matching criterion can be internal (e.g., observed total score or estimated latent variable score) or external (e.g., intelligence score, secondary test score). Furthermore, if it is internal, a decision has to be made about whether items that hold the potential for a lack of invariance should be included or excluded in the computation of the criterion because they may potentially contaminate its value; multiple procedures for purifying internal criteria exist.

An important distinction is made between bias and impact, uniform, nonuniform, and crossing lack of invariance, as well as item-, bundle-, and test-level lack of invariance. The first distinction between bias and impact allows one to distinguish between undesirable psychometric properties of the items (i.e., bias) and true average differences in abilities between the subgroups of interest (i.e., impact). This intuitively appealing and necessary distinction became more properly formalized in the research program of William Stout and colleagues, which is based on non-parametric IRT. The second distinction is best understood in the context of IRT or logistic regression analyses because smooth response curves can be considered. Here, uniform effects manifest themselves in parallel curves, nonuniform effects in curves that are

nonparallel but do not cross, and crossing effects in curves that do cross. The practical relevance of this is that, in the case of a set of curves that crosses, different groups are proportionally disadvantaged in different ranges of the score continuum, whereas otherwise, the effects always favor one or multiple groups over others. The third distinction between item- and test-level effects underscores that an item-level lack of invariance may be present that could either augment or vanish at higher levels, such as bundles or the test. This can result in the fact that effects of differing magnitudes are detected at the item level but may not affect practical decision making when sum scores are used if they cancel out; in practice, different levels of effects should be investigated.

Because of the multitude of methodological choices that are available to the practitioner, it can be difficult to synthesize results from various DIF studies. Moreover, the current literature has augmented the demands on researchers to go beyond empirically and rationally justifiable lack-of-invariance detection to provide potential explanations for its causes. In this respect, the use of variables that operationalize components of cognitive response processes to items may prove fruitful but are just beginning to be joined with lack-of-invariance analyses. Their use, along with the use of other noncognitive predictor variables, can be integrated eloquently into hierarchical logistic regression models.

If one integrates cognitive predictor variables into cognitive psychometric models, these parameters are also subject to an investigation of invariance. For some models, such as the DINA or the NIDA model, the linear logistic test model, as well as the rule-space methodology, these parameters are provided a priori by specialists so that the issue of parameter invariance can be resolved primarily theoretically. For other models, such as the reparameterized unified model, decisions about parameter inclusion and exclusion are made empirically so that the issue of parameter invariance should now be resolved through a synthesis of theoretical and empirical findings. How this can be accomplished in practice and whether a parameter invariance perspective is at all called for, is not yet agreed upon at this time.

Invariance Across Time Points

Less common, but equally important for large-scale testing companies in particular, is an investigation of parameter invariance over time. Due to the fact that a separate time point can be viewed abstractly as a mechanism that induces another grouping structure, many of the methods in the previous section for DIF, DBF, or DTF can be applied for this purpose with sometimes only minor modifications or, perhaps, even no modifications at all. However, the assessment of a lack of invariance in longitudinal models gets more complex when the models themselves get more complex. For example, in nonlinear growth models, some assumptions about a lack of invariance manifest themselves in very subtle parameter differences and, in some instances, may not be detectable with current software programs because of identification problems.

Conclusions

In conclusion, a lack of parameter invariance is somewhat of a paradox. It is simple to understand conceptually, but challenging to assess practically. Furthermore, even though methods for its assessment are either direct or slightly modified applications of existing statistical routines—with the exception of a few selected cases—understanding its nature, its directionality and the magnitude of its effects, and its potential causes is a challenging endeavor. But it is an indispensable endeavor to ascertain, from both empirical and rational perspectives, the validity of inferences about examinees and assessment instruments and the degree to which they are generalizable across measurement conditions.

—André A. Rupp

Further Reading

- De la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333–353.
- Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22, 33–51.

- Engelhard, G., Jr. (1994). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 73–99). Norwood, NJ: Ablex.
- Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education*, 11, 159–177.
- Hartz, S. M. (2002). *A Bayesian guide for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Seminar Press.
- McDonald, R. P. (1970). The theoretical foundations of common factor analysis, principal factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1–21.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Rupp, A. A. (2005). *A framework for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. Manuscript submitted for publication.
- Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, 21, 195–213.
- Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

PART AND PARTIAL CORRELATIONS

The semi-partial regression coefficient—also called part correlation—is used to express the specific portion of variance explained by a given independent variable in a multiple linear regression (MLR) analysis. It can be obtained as the correlation between the dependent variable and the residual of the prediction of one independent variable by the other ones. The semi-partial coefficient of correlation is used mainly in nonorthogonal multiple linear regression to assess the *specific* effect of each independent variable on the dependent variable.

The partial coefficient of correlation is designed to eliminate the effect of one variable on two other variables when assessing the correlation between these two variables. It can be computed as the correlation between the residuals of the prediction of these two variables by the first variable.

Multiple Regression Framework

In MLR, the goal is to predict, knowing the measurements collected on N subjects, a dependent variable Y from a set of K independent variables denoted

$$\{X_1, \dots, X_k, \dots, X_K\}. \quad (1)$$

We denote by $\mathbf{X}$ the $N \times (K + 1)$ augmented matrix collecting the data for the independent variables (this matrix is called augmented because the first column is composed only of ones), and by $\mathbf{y}$ the $N \times 1$ vector of observations for the dependent variable. These two matrices have the following structure:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} & \cdots & x_{1,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} & \cdots & x_{n,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,k} & \cdots & x_{N,K} \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix}. \quad (2)$$

The predicted values of the dependent variable $\hat{Y}$ are collected in a vector denoted $\hat{\mathbf{y}}$ and are obtained using MLR as

$$\mathbf{y} = \mathbf{X}\mathbf{b} \quad \text{with} \quad \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3)$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation, denoted $R^2_{Y,1,\dots,K}$. This coefficient is equal to the coefficient of correlation between the dependent variable (Y) and the predicted dependent variable ($\hat{Y}$).

Partial Regression Coefficient as Increment in Explained Variance

When the independent variables are pairwise orthogonal, the importance of each of them in the regression is assessed by computing the squared coefficient of

correlation between each of the independent variables and the dependent variable. The sum of these squared coefficients of correlation is equal to the squared multiple coefficient of correlation. When the independent variables are correlated, this strategy *overestimates* the contribution of each variable because the variance that they share is counted several times; and therefore, the sum of the squared coefficients of correlation is not equal to the multiple coefficient of correlation anymore. In order to assess the importance of a particular independent variable, the partial regression coefficient evaluates the *specific* proportion of variance explained by this independent variable. This is obtained by computing the increment in the multiple coefficient of correlation obtained when the independent variable is added to the other variables.

For example, consider the data given in Table 1, where the dependent variable is to be predicted from the independent variables X and T . The prediction equation (using Equation 3) is

$$\hat{Y} = 1.67 + X + 9.50T; \quad (4)$$

it gives a multiple coefficient of correlation of $R^2_{Y,XT} = .9866$. The coefficient of correlation between X and T is equal to $r_{X,T} = .7500$, between X and Y is equal to $r_{Y,X} = .8028$, and between T and Y is equal to $r_{Y,T} = .9890$. The squared partial regression coefficient between X and Y is computed as

$$r^2_{Y,X|T} = R^2_{Y,XT} - r^2_{Y,T} = .9866 - .9890^2 = .0085; \quad (5)$$

this indicates that when X is entered *last* in the regression equation, it increases the multiple coefficient of

correlation by .0085. In other words, X contributes a correlation of .0085 over and above the other dependent variable. As this example shows, the difference between the correlation and the part correlation can be very large. For T , we find that

$$r^2_{Y,T|X} = R^2_{Y,XT} - r^2_{Y,X} = .9866 - .8028^2 = .3421. \quad (6)$$

Partial Regression Coefficient as Prediction From a Residual

The partial regression coefficient can also be obtained by first computing for each independent variable the residual of its prediction from the other independent variables and then using this residual to predict the dependent variable. In order to do so, the first step is to isolate the specific part of each independent variable. This is done by first predicting a given independent variable from the other independent variables. The residual of the prediction is by definition uncorrelated with the predictors, hence, it represents the *specific* part of the independent variable under consideration.

We illustrate the procedure by showing how to compute the semi-partial coefficient between X and Y after the effect T has been partialled out. We denote by $\hat{X}_T$ the prediction of X from T .

The equation for predicting X from T is given by

$$\hat{X}_T = a_{X,T} + b_{X,T}T, \quad (7)$$

where $a_{X,T}$ and $b_{X,T}$ denote the intercept and slope of the regression line of the prediction of X from T .

Table 2 gives the values of the sums of squares and sum of cross-products needed to compute the prediction of X from T .

We find the following values for predicting X from T :

$$b_{X,T} = \frac{SCP_{XT}}{SS_T} = \frac{18}{16} = 1.125; \quad (8)$$

$$a_{X,T} = M_X - b_{X,T} \times M_T = 7 - 1.125 \times 3 = 3.625. \quad (9)$$

So, the first step is to predict one independent variable from the other one. Then, by subtracting the predicted value of the independent variable from its actual value, we obtain the residual of the prediction of this

Table 1 A Set of Data: Y Is to Be Predicted From X_1 and X_2

Y (Memory span)	14	23	30	50	39	67
X_1 (age)	4	4	7	7	10	10
X_2 (Speech rate)	1	2	2	4	3	6

Source: Data from Abdi et al. (2002).

Notes: Y is the number of digits a child can remember for a short time (the “memory span”), X_1 is the age of the child, and X_2 is the speech rate of the child (how many words the child can pronounce in a given time). Six children were tested.

Table 2 The Different Quantities Needed to Compute the Values of the Parameters $a_{X.T}$, $b_{X.T}$

X	x	x^2	T	t	t^2	$x \times t$	
4	-3	9	1	-2	4	6	
4	-3	9	2	-1	1	3	
7	0	0	2	-1	1	0	
7	0	0	4	1	1	0	
10	3	9	3	0	0	0	
10	3	9	6	3	9	9	
Σ	42	0	36	18	0	16	18
		SS_X		SS_T		SCP_{XT}	

Note: The following abbreviations are used: $x = (X - M_X)$,
 $t = (T - M_T)$.

independent variable. The residual of the prediction of X by T is denoted $e_{X.T}$ and it is computed as

$$e_{X.T} = X - \hat{X}_T. \quad (10)$$

Table 3 gives the quantities needed to compute $r_{Y.X|T}^2$. It is obtained as

$$r_{Y.X|T}^2 = r_{Y.e_{X.T}}^2 = \frac{(SCP_{Ye_{X.T}})^2}{SS_Y SS_{e_{X.T}}}. \quad (11)$$

In our example, we find

$$r_{Y.X|T}^2 = \frac{15.75^2}{1,846.83 \times 15.75} = .0085.$$

F and t Tests for the Partial Regression Coefficient

The partial regression coefficient can be tested by using a standard F test with the following degrees of freedom $v_1 = 1$ and $v_2 = N - K - 1$ (with N being the number of observations and K being the number of predictors). Because v_1 is equal to 1, the square root of F gives a Student t test. The computation of F is best described with an example: The F for the variable X in our example is obtained as

$$\begin{aligned} F_{Y.X|T} &= \frac{r_{Y.X|T}^2}{1 - R_{Y.X|T}^2} \times (N - 3) \\ &= \frac{.0085}{1 - .9866} \times 3 = 1.91. \end{aligned}$$

The relations between the partial regression coefficient and the different correlation coefficients are illustrated in Figure 1.

Alternative Formulas for the Semi-Partial Correlation Coefficients

The semi-partial coefficient of correlation can also be computed directly from the different coefficients of correlation of the independent variables and the dependent variable. Specifically, we find that the

Table 3 The Different Quantities to Compute the Semi-Partial Coefficient of Correlation Between Y and X After the Effects of T Have Been Partialled out of X

Y	y	y^2	X	$\hat{X}_T$	$e_{_{X.T}}$	$e^2_{_{X.T}}$	$y \times e_{_{X.T}}$
14	-23.1667	536.69	4	4.7500	-0.7500	0.5625	17.3750
23	-14.1667	200.69	4	5.8750	-1.8750	3.5156	26.5625
30	-7.1667	51.36	7	5.8750	1.1250	1.2656	-8.0625
50	12.8333	164.69	7	8.1250	-1.1250	1.2656	-14.4375
39	1.8333	3.36	10	7.0000	3.0000	9.0000	5.5000
67	29.8333	890.03	10	10.3750	-0.3750	0.1406	-11.1875
Σ	223	0	1,846.83	42	42.0000	0	15.7500
SS_Y						$SS_{e_{_{X.T}}}$	$SCP_{_{\hat{y}_{_{X.T}}}}$

Note: The following abbreviations are used: $y = Y - M_Y$, $e_{X.T} = X - \hat{X}_T$.

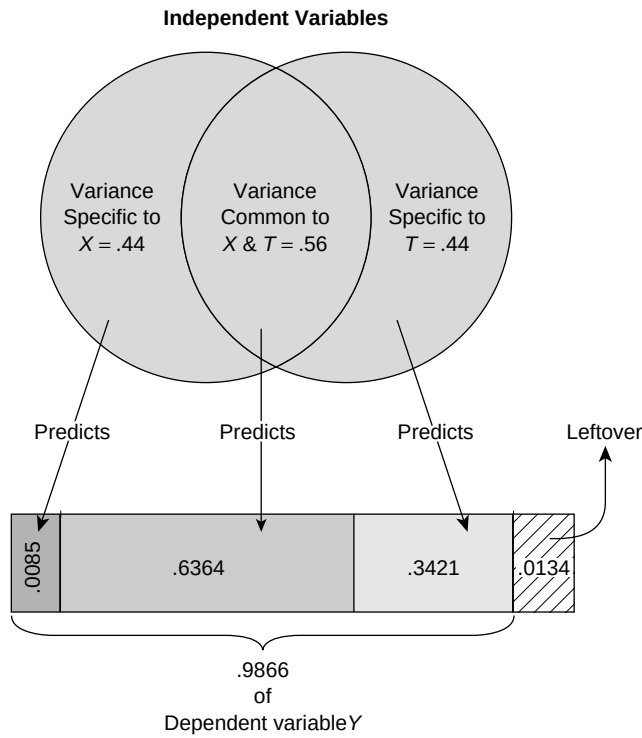


Figure 1 Illustration of the Relationship of the Independent Variables

Notes: The dependent variable shows which part of the independent variables explains what proportion of the dependent variable. The independent variables are represented by a Venn diagram, and the dependent variable is represented by a bar.

semi-partial correlation between Y and X can be computed as

$$r_{Y.X|T}^2 = \frac{(r_{Y.X} - r_{Y.T}r_{X.T})^2}{1 - r_{X.T}^2}. \quad (12)$$

For our example, taking into account that

$$r_{X.T} = .7500,$$

$$r_{YX} = .8028, \text{ and}$$

$$r_{YT} = .9890,$$

we find that

$$\begin{aligned} r_{Y.X|T}^2 &= \frac{(r_{Y.X} - r_{Y.T}r_{X.T})^2}{1 - r_{X.T}^2} \\ &= \frac{(.8028 - .9890 \times .7500)^2}{1 - .7500^2} \approx .0085. \end{aligned} \quad (13)$$

Partial Correlation

When dealing with a set of dependent variables, sometimes we want to evaluate the correlation between two dependent variables after the effect of a third dependent variable has been removed from *both* dependent variables. This can be obtained by computing the coefficient of correlation between the residuals of the prediction of each of the first two dependent variables by the third dependent variable (i.e., if you want to eliminate the effect of, say, variable Q from variables Y and W , you first predict Y from Q and W from Q , and then you compute the residuals and correlate them). This coefficient of correlation is called a *partial* coefficient of correlation. It can also be computed directly using a formula involving only the coefficients of correlation between pairs of variables. As an illustration, suppose that we want to compute the squared partial coefficient of correlation between Y and X after having eliminated the effect of T from both of them (this is done only for illustrative purposes because X and T are independent variables, not dependent variables). This coefficient is noted $r_{(Y.X)|T}^2$ (read “ r square of Y and X after T has been partialled out from Y and X ”), and it is computed as

$$r_{(Y.X)|T}^2 = \frac{(r_{Y.X} - r_{Y.T}r_{X.T})^2}{(1 - r_{Y.T}^2)(1 - r_{X.T}^2)}. \quad (14)$$

For our example, taking into account that

$$r_{X.T} = .7500,$$

$$r_{YX} = .8028, \text{ and}$$

$$r_{YT} = .9890,$$

we find the following values for the partial correlation of Y and X :

$$\begin{aligned} r_{(Y.X)|T}^2 &= \frac{(r_{Y.X} - r_{Y.T}r_{X.T})^2}{(1 - r_{Y.T}^2)(1 - r_{X.T}^2)} \\ &= \frac{(.8028 - .9890 \times .7500)^2}{(1 - .9890^2)(1 - .7500^2)} \approx .3894. \end{aligned} \quad (15)$$

Further Reading

Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition.

PARTIAL LEAST SQUARE REGRESSION

Partial least square (PLS) regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression. Its goal is to predict or analyze a set of dependent variables from a set of independent variables or predictors. This prediction is achieved by extracting from the predictors a set of orthogonal factors called *latent variables*, which have the best predictive power.

PLS regression is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences (specifically economics) with Herman Wold, but became popular first in chemometrics (i.e., computational chemistry), due in part to Herman's son Svante, and in sensory evaluation. But PLS regression is also becoming a tool of choice in the social sciences as a multivariate technique for nonexperimental and experimental data alike (e.g., neuroimaging). It was first presented as an algorithm akin to the power method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework.

Prerequisite Notions and Notations

The I observations described by K dependent variables are stored in a $I \times K$ matrix denoted $\mathbf{Y}$, and the values of J predictors collected on these I observations are collected in the $I \times J$ matrix $\mathbf{X}$.

Goal

The goal of PLS regression is to predict $\mathbf{Y}$ from $\mathbf{X}$ and to describe their common structure. When $\mathbf{Y}$ is a vector and $\mathbf{X}$ is full rank, this goal could be accomplished using ordinary multiple regression. When the number

of predictors is large compared to the number of observations, $\mathbf{X}$ is likely to be singular and the regression approach is no longer feasible (i.e., because of multicollinearity). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using stepwise methods); another one, called principal component regression, is to perform a principal component analysis (PCA) of the $\mathbf{X}$ matrix and then use the principal components (i.e., eigenvectors) of $\mathbf{X}$ as regressors on $\mathbf{Y}$. Technically, in PCA, $\mathbf{X}$ is decomposed using its singular value decomposition as $\mathbf{X} = \mathbf{S}\mathbf{\Delta}\mathbf{V}^T$ with $\mathbf{S}^T\mathbf{S} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, and $\mathbf{\Delta}$ being a diagonal matrix with the singular values as diagonal elements. The singular vectors are ordered according to their corresponding singular values, which correspond to the square root of the variance of $\mathbf{X}$ explained by each singular vector. The left singular vectors (i.e., the columns of $\mathbf{S}$) are then used to predict $\mathbf{Y}$ using standard regression because the orthogonality of the singular vectors eliminates the multicollinearity problem. But the problem of choosing an *optimum* subset of predictors remains. A possible strategy is to keep only a few of the first components. But these components are chosen to explain $\mathbf{X}$ rather than $\mathbf{Y}$, and so, nothing guarantees that the principal components, which “explain” $\mathbf{X}$, are relevant for $\mathbf{Y}$.

By contrast, PLS regression finds components from $\mathbf{X}$ that are also relevant for $\mathbf{Y}$. Specifically, PLS regression searches for a set of components (called *latent vectors*) that performs a simultaneous decomposition of $\mathbf{X}$ and $\mathbf{Y}$ with the constraint that these components explain as much as possible of the *covariance* between $\mathbf{X}$ and $\mathbf{Y}$. This step generalizes PCA. It is followed by a regression step where the decomposition of $\mathbf{X}$ is used to predict $\mathbf{Y}$.

Simultaneous Decomposition of Predictors and Dependent Variables

PLS regression decomposes both $\mathbf{X}$ and $\mathbf{Y}$ as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are *decomposed* as $\mathbf{X} = \mathbf{T}\mathbf{P}^T$ with $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, with $\mathbf{I}$ being the identity matrix (some variations of the technique do

not require $\mathbf{T}$ to have unit norms). By analogy, with PCA, $\mathbf{T}$ is called the *score* matrix and $\mathbf{P}$ the *loading* matrix (in PLS regression, the loadings are not orthogonal). Likewise, $\mathbf{Y}$ is *estimated* as $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T$, where $\mathbf{B}$ is a diagonal matrix with the “regression weights” as diagonal elements and $\mathbf{C}$ is weight matrices for the dependent variables (see below for more details on these two matrices). The columns of $\mathbf{T}$ are the *latent vectors*. When their number is equal to the rank of $\mathbf{X}$, they perform an exact decomposition of $\mathbf{X}$. Note, however, that they only *estimate* $\mathbf{Y}$ (i.e., in general, $\hat{\mathbf{Y}}$ is not equal to $\mathbf{Y}$).

PLS Regression and Covariance

The latent vectors could be chosen in a lot of different ways. In fact, in the previous formulation, any set of orthogonal vectors spanning the column space of $\mathbf{X}$ could be used to play the role of $\mathbf{T}$. In order to specify $\mathbf{T}$, additional conditions are required. For PLS regression, this amounts to finding two sets of weights $\mathbf{w}$ and $\mathbf{c}$ in order to create (respectively) a linear combination of the columns of $\mathbf{X}$ and $\mathbf{Y}$ such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{c}$ with the constraints that $\mathbf{w}^T\mathbf{w} = 1$, $\mathbf{t}^T\mathbf{t} = 1$, and $\mathbf{t}^T\mathbf{u}$ is maximal. When the first latent vector is found, it is *subtracted* from both $\mathbf{X}$ and $\mathbf{Y}$, and the procedure is reiterated until $\mathbf{X}$ becomes a null matrix (see the algorithm section for more).

A PLS Regression Algorithm

The properties of PLS regression can be analyzed from a sketch of the original algorithm. The first step is to create two matrices: $\mathbf{E} = \mathbf{X}$ and $\mathbf{F} = \mathbf{Y}$. These matrices are then column centered and normalized (i.e., transformed into z scores). The sum of squares of these matrices are denoted SS_X and SS_Y . Before starting the iteration process, the vector $\mathbf{u}$ is initialized with random values (in what follows, the symbol $\propto$ means “to normalize the result of the operation”).

Step 1. $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$ (estimate $\mathbf{X}$ weights)

Step 2. $\mathbf{t} \propto \mathbf{E}\mathbf{w}$ (estimate $\mathbf{X}$ factor scores)

Step 3. $\mathbf{c} \propto \mathbf{F}^T\mathbf{t}$ (estimate $\mathbf{Y}$ weights)

Step 4. $\mathbf{u} = \mathbf{F}\mathbf{c}$ (estimate $\mathbf{Y}$ scores)

If $\mathbf{t}$ has not converged, then go to Step 1; if $\mathbf{t}$ has converged, then compute the value of b , which is used to predict $\mathbf{Y}$ from $\mathbf{t}$ as $b = \mathbf{t}^T\mathbf{u}$, and compute the factor loadings for $\mathbf{X}$ as $\mathbf{p} = \mathbf{E}^T\mathbf{t}$. Now subtract (i.e., partial out) the effect of $\mathbf{t}$ from both $\mathbf{E}$ and $\mathbf{F}$ as follows: $\mathbf{E} = \mathbf{E} - \mathbf{t}\mathbf{p}^T$ and $\mathbf{F} = \mathbf{F} - b\mathbf{t}\mathbf{c}^T$. The vectors $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{c}$, and $\mathbf{p}$ are then stored in the corresponding matrices, and the scalar b is stored as a diagonal element of $\mathbf{B}$. The sum of squares of $\mathbf{X}$ (respectively, $\mathbf{Y}$) explained by the latent vector is computed as $\mathbf{p}^T\mathbf{p}$ (respectively, b^2), and the proportion of variance explained is obtained by dividing the explained sum of squares by the corresponding total sum of squares (i.e., SS_X and SS_Y).

If $\mathbf{E}$ is a null matrix, then the whole set of latent vectors has been found; otherwise, the procedure can be reiterated from Step 1 on.

PLS Regression and the Singular Value Decomposition

The iterative algorithm presented above is similar to the power method, which finds eigenvectors. So PLS regression is likely to be closely related to the eigen and singular value decompositions, and this is indeed the case. For example, if we start from Step 1, which computes $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$, and substitute the rightmost term iteratively, we find the following series of equations: $\mathbf{w} \propto \mathbf{E}^T\mathbf{u} \propto \mathbf{E}^T\mathbf{F}\mathbf{c} \propto \mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{t} \propto \mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{E}\mathbf{w}$. This shows that the first weight vector $\mathbf{w}$ is the first right singular vector of the matrix $\mathbf{X}^T\mathbf{Y}$. Similarly, the first weight vector $\mathbf{c}$ is the left singular vector of $\mathbf{X}^T\mathbf{Y}$. The same argument shows that the first vectors $\mathbf{t}$ and $\mathbf{u}$ are the first eigenvectors of $\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{X}^T$.

Prediction of the Dependent Variables

The dependent variables are predicted using the multivariate regression formula as $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T = \mathbf{X}\mathbf{B}_{\text{PLS}}$ with $\mathbf{B}_{\text{PLS}} = (\mathbf{P}^{T+})\mathbf{B}\mathbf{C}^T$ (where $\mathbf{P}^{T+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{P}^T$). If all the latent variables of $\mathbf{X}$ are used, this regression is equivalent to principal

component regression. When only a subset of the latent variables is used, the prediction of $\mathbf{Y}$ is optimal for this number of predictors.

An obvious question is to find the number of latent variables needed to obtain the best generalization for the prediction of *new* observations. This is, in general, achieved by cross-validation techniques such as bootstrapping.

The interpretation of the latent variables is often helped by examining graphs akin to PCA graphs (e.g., by plotting observations in a $\mathbf{t}_1 \times \mathbf{t}_2$ space; see Figure 1).

A Small Example

We want to predict the subjective evaluation of a set of five wines. The dependent variables that we want to predict for each wine are its likeability and how well it goes with meat or dessert (as rated by a panel of experts) (see Table 1). The predictors are the price and the sugar, alcohol, and acidity content of each wine (see Table 2).

The different matrices created by PLS regression are given in Tables 3 to 11. From Table 11, one can

Table 1 The $\mathbf{Y}$ Matrix of Dependent Variables

Wine	Hedonic	Goes With Meat	Goes With Dessert
1	14	7	8
2	10	7	6
3	8	5	5
4	2	4	7
5	6	2	4

Table 2 The $\mathbf{X}$ Matrix of Predictors

Wine	Price	Sugar	Alcohol	Acidity
1	7	7	13	7
2	4	3	14	7
3	10	5	12	5
4	16	7	11	3
5	13	3	10	3

find that two latent vectors explain 98% of the variance of $\mathbf{X}$ and 85% of $\mathbf{Y}$. This suggests keeping these two dimensions for the final solution. The examination of the two-dimensional regression coefficients

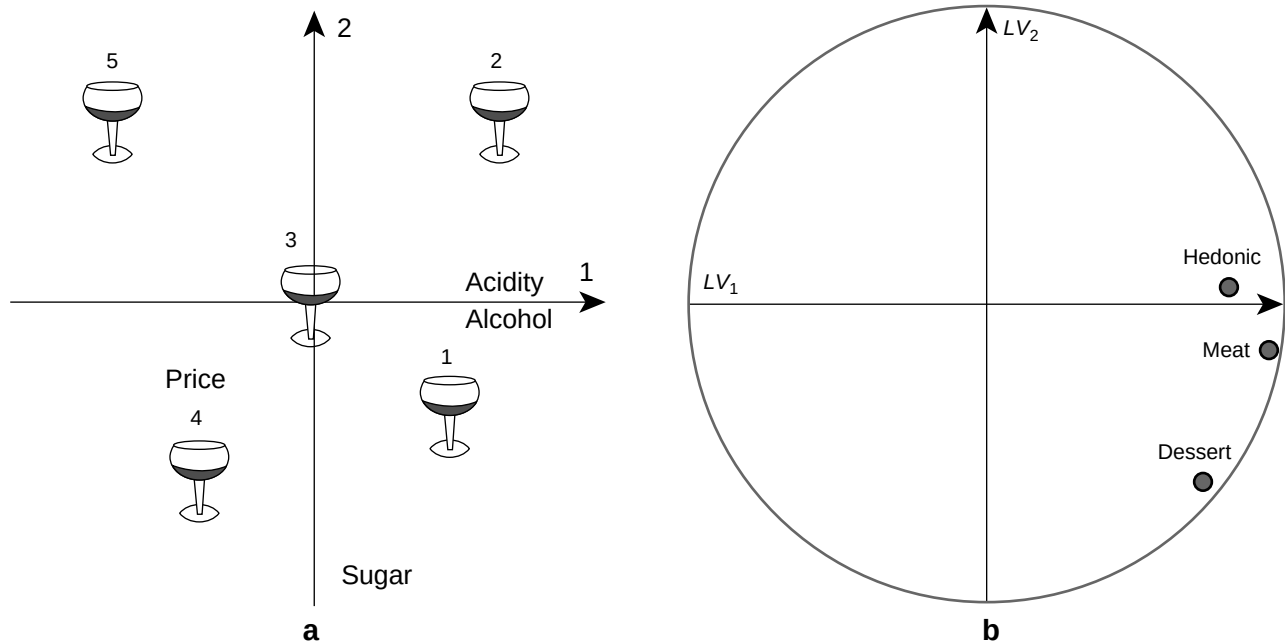


Figure 1 PLS Regression

Note: (a) Projection of the wines and the predictors on the first two latent vectors (respectively, matrices $\mathbf{T}$ and $\mathbf{W}$); (b) circle of correlation showing the correlation between the original dependent variables (matrix $\mathbf{Y}$) and the latent vectors (matrix $\mathbf{T}$).

Table 3 The Matrix **T**

<i>Wine</i>	t_1	t_2	t_3
1	0.4538	-0.4662	0.5716
2	0.5399	0.4940	-0.4631
3	0	0	0
4	-0.4304	-0.5327	-0.5301
5	-0.5633	0.5049	0.4217

Table 4 The Matrix **U**

<i>Wine</i>	u_1	u_2	u_3
1	1.9451	-0.7611	0.6191
2	0.9347	0.5305	-0.5388
3	-0.2327	0.6084	0.0823
4	-0.9158	-1.1575	-0.6139
5	-1.7313	0.7797	0.4513

Table 5 The Matrix **P**

	p_1	p_2	p_3
Price	-1.8706	-0.6845	-0.1796
Sugar	0.0468	-1.9977	0.0829
Alcohol	1.9547	0.0283	-0.4224
Acidity	1.9874	0.0556	0.2170

Table 6 The Matrix **W**

	w_1	w_2	w_3
Price	-0.5137	-0.3379	-0.3492
Sugar	0.2010	-0.9400	0.1612
Alcohol	0.5705	-0.0188	-0.8211
Acidity	0.6085	0.0429	0.4218

Table 7 The Matrix **B**_{PLS} When Three Latent Vectors Are Used

	<i>Hedonic</i>	<i>Goes With Meat</i>	<i>Goes With Dessert</i>
Price	-1.0607	-0.0745	0.1250
Sugar	0.3354	0.2593	0.7510
Alcohol	-1.4142	0.7454	0.5000
Acidity	1.2298	0.1650	0.1186

Table 8 The Matrix **B**_{PLS} When Two Latent Vectors Are Used

	<i>Hedonic</i>	<i>Goes With Meat</i>	<i>Goes With Dessert</i>
Price	-0.2662	-0.2498	0.0121
Sugar	0.0616	0.3197	0.7900
Alcohol	0.2969	0.3679	0.2568
Acidity	0.3011	0.3699	0.2506

Table 9 The Matrix **C**

	c_1	c_2	c_3
Hedonic	0.6093	0.0518	0.9672
Goes with meat	0.7024	-0.2684	-0.2181
Goes with dessert	0.3680	-0.9619	-0.1301

Table 10 The **b** Vector

b_1	b_2	b_3
2.7568	1.6272	1.1191

Table 11 Variance of **X** and **Y** Explained by the Latent Vectors

<i>Latent Vector</i>	<i>Percentage of Explained Variance for X</i>	<i>Cumulative Percentage of Explained Variance for X</i>	<i>Percentage of Explained Variance for Y</i>	<i>Cumulative Percentage of Explained Variance for Y</i>
1	70	70	63	63
2	28	98	22	85
3	2	100	10	95

(i.e., $\mathbf{B}_{\text{PLS}}$, see Table 8) shows that sugar is mainly responsible for choosing a dessert wine, and that price is negatively correlated with the perceived quality of the wine, whereas alcohol is positively correlated with it (at least in this example). Looking at the latent vectors shows that $\mathbf{t}_1$ expresses price and $\mathbf{t}_2$ reflects sugar content. This interpretation is confirmed and illustrated in Figure 1a, which displays the projections on the latent vectors of the wines (matrix $\mathbf{T}$) and the predictors (matrix $\mathbf{W}$), and in Figure 1b, which displays the correlation between the original dependent variables and the projection of the wines on the latent vectors.

Relationship With Other Techniques

PLS regression is obviously related to canonical correlation, STATIS, and to multiple factor analysis. The main originality of PLS regression is to preserve the asymmetry of the relationship between predictors and dependent variables, whereas these other techniques treat them symmetrically.

Software

PLS regression necessitates sophisticated computations, and therefore, its application depends on the availability of software. For chemistry, two main programs are used: the first one, called SIMCA-P, was developed originally by Wold; the second one, called the Unscrambler, was first developed by Martens, who was another pioneer in the field. For brain imaging, SPM, which is one of the most widely used programs in this field, has recently integrated a PLS regression module. Outside these domains, SAS PROC PLS is probably the most easily available program. In addition, interested readers can download a set of MATLAB programs from the author's home page (www.utdallas.edu/~herve). Also, a public domain set of MATLAB programs is available from the home page of the N-Way project (www.models.kvl.dk/source/nwaytoolbox/), along with tutorials and examples. From brain imaging, a special toolbox written in MATLAB is freely available from www.rotman-baycrest.on.ca. And finally, a commercial MATLAB toolbox has also been developed by Eigenresearch.

—Hervé Abdi

See also DISTATIS; Eigendecomposition; Multicollinearity; Multiple Correspondence Analysis; Multiple Factor Analysis; Principal Component Analysis; Regression Analysis; R_v and Congruence Coefficients; Singular and Generalized Singular Value Decomposition; STATIS; Structural Equation Modeling; z Scores

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H. (2004). PLS-regression. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
- Martens, H., & Naes, T. (1989). *Multivariate calibration*. London: Wiley.
- McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage*, 23, 250–263.
- Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and intelligent laboratory systems*, 58, 261–273.
- Phatak, A., & de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, 11, 311–338.
- Tenenhaus, M. (1998). *La régression PLS*. Paris: Technip.
- Ter Braak, C. J. F., & de Jong, S. (1998). The objective function of partial least squares regression. *Journal of Chemometrics*, 12, 41–54.
- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.
- Worsley, K. J. (1997). An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5, 254–258.

PASCAL, BLAISE (1623–1662)

Pascal was born at Clermont-Ferrand in France on June 19, 1623. His mother died when he was 3 years old. He was an obviously bright child, and his father initially forbade his study of mathematics, preferring that he become fluent in the classics instead. Of course, this

merely spurred Pascal's curiosity, and he swiftly began solving geometric problems for himself. By the age of 17, he had published a paper (on conic sections), and at 18, he invented his first calculator. This was the first of several adding machines, some of which cleverly worked to different bases (e.g. both 12 and 20).

From mathematics, Pascal turned to physics, observing that the atmospheric pressure was lower at the top of the mountain than at the bottom. This observation had followed a visit by Descartes, in which the two had discussed whether a vacuum could exist. Apparently, Descartes left, remarking that Pascal "had too much vacuum in his head."

When Pascal's father injured his leg, religious brothers cared for Pascal, and this marked the beginnings of Pascal's religious conversion. In 1650, he took a break from science to consider religion, but following the death of his father in 1653, he returned to scientific and mathematical studies. At this time, he introduced the numerical configuration now called Pascal's triangle. His correspondence with Fermat laid the foundations for the modern treatment of probability.

Once again, however, fate took a hand. In 1654, he had a traffic accident: his carriage veered off the road, and he escaped death only because the harness straps broke. Shortly afterwards, Pascal had a vision: He wrote an account of this vision that he kept on his person for the rest of his life. He then began writing philosophical and theological pieces. The last and best known of these was entitled *Pensées* (Thoughts), which was unfinished at the time of his premature death in Paris on August 19, 1662. It is in the *Pensées* that Pascal argued that it is always a better "bet" to believe in God than not to believe in God. This is because the expected value to be gained from believing in God must be greater than the expected value for nonbelievers!

—Graham Upton

Further Reading

Rogers, B. (1998). *Pascal: In praise of vanity*. London: Phoenix.

A Short Account of the History of Mathematics (4th ed., 1908) by W. W. Rouse Ball [excerpt]: http://www.maths.tcd.ie/pub/HistMath/People/Pascal/RouseBall/RB_Pascal.html

PATH ANALYSIS

Path analysis is a method of describing and quantifying the relationship of a dependent variable Y to a set of other variables. Each of the variables may have a direct effect on Y , as well as indirect effects via other variables. It is difficult to infer causation without an experiment, but by describing a system of interrelationships, path analysis can take a step to support cause-and-effect inference.

Path analysis was originally developed by Sewall Wright for the field of genetics. It has since been adopted by virtually all the behavioral sciences and applied to a variety of research questions. Properly applied, it can be an aid to scientific, business, governmental, and educational policy decisions.

Examples

One well-known example relates income to occupation, education, father's education, father's occupation, and number of siblings. In another example, a professor wants to relate students' scores on the final exam to those on the two exams during the term. Exam 1 can be related to the final exam score either directly or indirectly via Exam 2. What are the sizes of these effects of Exam 1? If both are small, then the professor might give only a midterm and a final next semester. In a third example, Profit can be studied as a function of Sales and Assets. Assets support Sales, and hence indirectly contribute to Profit. But Assets can also contribute directly to Profit, via interest and rents. What are the relative sizes of these effects? (See Table 1.)

In considering the relationship between Profit, Sales, and Assets, it becomes apparent that Assets, by supporting Sales, contribute indirectly to Profit. Assets also contribute directly to Profit via dividends on investments, interest on savings, and rents on property. The prediction equation for Profit based on Sales and Assets, estimated by the statistical technique of multiple linear regression, is

$$\text{predicted Profit} = 1464 + 0.010348 \text{ Sales} + 0.010069 \text{ Assets (\$M)}.$$

Table 1 Sales, Profits, and Assets of the 10 Largest U.S. Industrial Corporations (in millions of dollars)

Company	Sales	Profits	Assets
GM	126974	4224	173297
Ford	96933	3835	160893
Exxon	86656	3510	83219
IBM	63438	3758	77734
GE	55264	3939	128344
Mobil	50976	1809	39080
Philip Morris	39069	2946	38528
Chrysler	36156	359	51038
Du Pont	35209	2480	34715
Texaco	32416	2413	25636

Source: "Fortune 500," *Fortune*, 121 (April 23, 1990), 346–367.
© 1990 Time Inc. All rights reserved.

The fitting of such a relationship is called *regressing* Profit on Sales and Assets. The interpretation of the coefficient (multiplier) of Assets is that if Assets increase by one unit (1 \$M) and Sales remain the same, Profit is predicted to increase by 0.010069. But if Assets increase by one unit, what are Sales expected to do? Would not higher Assets possibly mean higher Sales? We need also to regress Sales on Assets. This gives

$$\text{predicted Sales} = 21097 + 0.507 \text{ Assets } (\$M).$$

So, in view of our understanding of Assets as contributing both directly and indirectly to Profit, we really need this system of two regression equations, one for Sales on Assets, the other for Profit on Sales and Assets. Now we can see that if Assets increase by 1 unit, Sales are expected to increase by 0.507, and the expected change in Profit is

$$0.010348(0.507) + 0.010069(1) = 0.015315 (\$M).$$

The system can be described by the paths

$$\text{Assets} \rightarrow \text{Sales} \rightarrow \text{Profit}, \text{Assets} \rightarrow \text{Profit};$$

The variable Assets has a direct connection to Profit and also an indirect connection via Sales.

Regression

In more general terms, path analysis employs a system of *regression* equations. Regression is a method of studying the dependence of a variable Y on several explanatory variables $X_1, X_2, \dots, X_p$. The data set is of the form

$$\{(x_{1i}, x_{2i}, \dots, x_{pi}, y_i), i = 1, 2, \dots, n \text{ cases}\}.$$

A prediction equation of the form

$$Y' = a + b_1X_1 + b_2X_2 + \dots + b_pX_p,$$

where Y' is the predicted value of Y , is obtained. The coefficients $a, b_1, \dots, b_p$ are obtained from the data set by a method such as least squares. Values of the X s for a new individual can be substituted into the equation to predict the value of Y for that individual. But regression usually raises more questions than it answers. It does not deal explicitly with relationships among the X s. The coefficient of an X gives the predicted change in Y corresponding to a unit change in that X , given that the other X s remain constant. But usually if the value of one X changes, some of the other X s, correlated with it, will change too.

A System of Relationships

In path analysis, variables explained in terms of others are called *endogenous* and denoted by Y s; variables not predicted by others are called *exogenous* and denoted by X s. Causation can be illustrated by paths. Even in a system of one dependent variable and two other variables, there are several possibilities, including the following:

$$X \rightarrow Y_1 \rightarrow Y_2$$

This path indicates that X affects Y_1 , and Y_1 affects Y_2 . An alternative model is the following system of two paths:

$$X \rightarrow Y_1 \rightarrow Y_2, X \rightarrow Y_2$$

This indicates that X affects Y_1 , and Y_1 affects Y_2 , but also X has a direct effect of its own on Y_2 .

From a postulated model, specifying relationships among the variables, one can derive estimates of the *path coefficients*, the weights on the arcs of the paths. These are obtained by estimating the system of equations corresponding to the paths. Each endogenous variable is regressed on all the variables pointing directly to it in paths.

For the model consisting of a single path, there would be two simultaneous equations, one showing dependence of Y_2 on Y_1 ; the other, that of Y_1 on X :

$$[1]: Y_2' = a + bY_1$$

$$[2]: Y_1' = c + dX$$

For the model consisting of two paths, in which X has both a direct and an indirect effect on Y_2 , there would again be two equations, but now one shows dependence of Y_2 on both Y_1 and X :

$$[1]: Y_2' = a + bX + cY_1$$

$$[2]: Y_1' = d + eX$$

The input for path analysis can be the set of correlations between the pairs of variables. Use of correlations is equivalent to standardizing the data to z scores at the outset. The z score is the raw score, minus the mean for the variable, divided by the standard deviation for the variable.

Assumptions Underlying Path Analysis

The sample size needs to be adequate; guidelines for the minimum range from at least 5 cases per variable to at least 10 cases per variable. (The sample size in the Assets-Sales-Profit example is purposefully small.) The usual assumptions of linearity and constant variance (variance not dependent upon the values of the explanatory variables) underlying the use of multiple linear regression apply to each of the regressions in the model. For use of the t distribution for hypothesis tests, the errors should be at least

approximately normally distributed. Path analysis is not without its detractors, at least in part because of the difficulty of checking these assumptions for a whole set of equations.

Related Techniques

The variables in path analysis are observable (*manifest*). Structural equation modeling (SEM) includes models with hidden (*latent*) variables. In SEM, concepts (*constructs*) are linked via the structural model. Each construct is entered via its measurement model, involving a factor analysis of several indicators, which are observable variables that relate to it. The factor analysis produces a combination of the indicators for that construct.

Example

Sears, Roebuck & Co. used SEM in a new employee-customer-profit corporate model developed in the early 1990s. The constructs were employee satisfaction, customer satisfaction, and profit. Each construct was measured by several indicators; for example, profit was measured by the customary accounting variables. The indicators for each construct are combined by factor analysis into an index of that construct. Profit in each fiscal quarter was predicted from customer satisfaction in the preceding quarter, which in turn was related to employee satisfaction in the quarter before that.

Path Diagrams

The paths for a larger model, such as the first example given above, look like this.

Let FE be father's education, FO be father's occupation, S be number of siblings, E be education, O be occupation, and I be income. These paths might be those considered.

$$FE \rightarrow FO \rightarrow S \rightarrow E \rightarrow O \rightarrow I$$

$$FE \rightarrow FO \rightarrow S \rightarrow E \rightarrow O$$

$$FE \rightarrow FO \rightarrow S$$

A *path diagram* is a way of drawing the system. The diagram, made up of paths like the ones above, is constructed as follows:

1. A straight arrow is drawn to each endogenous variable from each of its sources.
2. A straight arrow is drawn to each dependent variable from its residual. Each endogenous variable has a residual because it is the dependent variable in a regression.
3. A curved, double-headed arrow is drawn between each pair of exogenous variables thought to have nonzero correlation.

Conditions for Estimability

To estimate the path coefficients, equations are set up between the path coefficients and the variances and covariances of the p endogenous and q exogenous variables. The complexity of the model considered is limited by the number of variables in the data set. It is necessary that the number t of path-model parameters be no more than the number $(p + q)(p + q + 1)/2$ of variances and covariances of the $p + q$ variables. If this condition is satisfied, then, by a process analogous to checking whether a system of linear equations consists of independent equations, it is determined whether the system is solvable.

Comparing Alternative Models

Note that path analysis and SEM are mainly concerned with estimating the sizes of effects in an already postulated model. However, sometimes, there is a search among a number of models. Members of a research team may propose competing models. Path diagrams provide a way of brainstorming to organize a research project into the exploration of competing models. Then there is the problem of choosing among them.

Model-selection criteria are figures of merit for scoring alternative models. Prominent among these are criteria equal to a measure of lack of fit of the model (analogous to residual sum of squares in

regression), plus a multiple of the number of parameters used to achieve that fit. The smaller the score on such a criterion, the better the model.

Using the Computer

Although there was path analysis software before graphical user interfaces, these can be used now to draw and modify diagrams. Given the diagram, the software checks estimability of the model and then fits the model. Then, the diagram can be displayed with the path coefficients on it. Software for path analysis and SEM includes AMOS, EQS, and LISREL.

—Stanley L. Sclove

Further Reading

- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.
- Duncan, O. D., & Hodge, R. W. (1963). Education and occupational mobility: A regression analysis. *American Journal of Sociology*, 68, 629–644.
- Li, C. C. (1975). *Path analysis: A primer*. Pacific Grove, CA: Boxwood Press.
- Rucci, A. J., Kim, S. P., & Quinn, R. T. (1998, January-February). The employee-customer-profit chain at Sears. *Harvard Business Review*, pp. 82–97.

Path diagrams and software for path analysis and SEM:
<http://www.spss.com/amos/>

PEABODY PICTURE VOCABULARY TEST

The Peabody Picture Vocabulary Test, now in its third edition (PPVT-III), is an individually administered, untimed, norm-referenced, wide-range test of listening comprehension for the spoken word in standard English. The test serves as both an achievement test of receptive vocabulary attainment for standard English

and a screening test of verbal ability, but only when English is the language of the examinee's home, community, and school. The test is designed for use with persons aged 2½ through 90+ years, and the average administration time is 10–15 minutes. Two types of normative scores are provided. Deviation-type norms tell you how far the individual's performance deviates from the average person of the same age group on whom the test was standardized. Developmental-type norms tell you how the individual's performance compares to others on a growth or development curve.

Two forms are available, Form IIIA and Form IIIB, that allow for reliable testing and retesting. Each form contains four training items and 204 test items. Items are grouped into 17 sets of 12 items, arranged in order of increasing difficulty. Each item contains four black-and-white illustrations arranged on a page. The examinee selects the picture that best represents the meaning of a stimulus word presented orally by the examiner. A sample item for the stimulus word *carpenter* could contain a picture of a carpenter with three distracter pictures of various workers such as a doctor, a chef, and a farmer.

The PPVT-III is useful with diverse populations. Reading or writing are not required from the examinee; therefore, it is useful with nonreaders and people with written-language problems. The examinee can respond orally, by pointing, or by signaling “yes” or “no” as the examiner points to each choice in turn. Because an oral, written, or point response is not required from the examinee, the test may be used with people with language impairments, and disabilities such as autism or cerebral palsy. The test can aid in the detection of language impairments, such as aphasia. The PPVT-III can be useful as a screening measure for both giftedness and mental retardation for people with English-language backgrounds. When English is a second language, the test can provide a measure of English-language proficiency. The ease of administration and high reliability at young ages makes the test useful in testing preschool children's vocabulary acquisition. The extensive adult norms make the test useful in testing adults'

listening comprehension and possible vocabulary deterioration.

—Ann M. Weber

Further Reading

- Campbell, J. (1998). Test reviews. *Journal of Psychoeducational Assessment*, 16, 334–338.
- Campbell, J., Bell, S., & Keith, L. (2001). Concurrent validity of the Peabody Picture Vocabulary Test-Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment*, 8(1), 85–94.
- Pankratz, M., Morrison, A., & Plante, E. (2004). Difference in standard scores of adults on the Peabody Picture Vocabulary Test (Revised and Third Edition). *Journal of Speech, Language, & Hearing Research*, 47(3), 714–718.
- Peabody Picture Vocabulary Test authors Lloyd M. Dunn and Leota M. Dunn biographies: <http://www.slpforum.com/bio/dunn.asp>

PEARSON, KARL (1857–1936)

Karl Pearson was judged by his contemporaries to be among the most influential university teachers of his time. The modern statistician sees him as a founding father of what are now considered to be elementary statistical methods.

Pearson, the son of a barrister, was born in London, on March 27, 1857. In 1875, he gained a scholarship to study mathematics at King's College, Cambridge. A somewhat rebellious student, Pearson spent much of his time studying other subjects. His wide range of interests was further extended when he spent time studying physics, metaphysics, and German literature at the Universities of Heidelberg and Berlin. It was in Germany that he changed his first name from Carl to Karl, retaining that spelling thereafter.

Returning to England, Pearson rejected the offer of a post in the German Department at Cambridge University, opting instead to study law. He was called to the Bar in 1882, but never practiced law. Instead, in 1885, he was appointed Goldsmid Professor in the

Applied Mathematics Department at University College, London.

In 1890, Pearson's career changed direction once again, with the appointment of Raphael Weldon as Professor of Zoology. Pearson's interaction with Weldon was stimulated by the publication of Francis Galton's *Natural Inheritance*. By the time of his first statistical publication (*The Chances of Death and Other Studies in Evolution*), he already had 100 publications to his name (including many on German history and folklore). However, between 1893 and 1912, Pearson wrote 18 papers under the general title of *Mathematical Contributions to the Theory of Evolution*. These papers resulted in the introduction of such familiar concepts as standard deviation, regression, correlation, and the chi-square goodness-of-fit test.

In 1901, Pearson was a co-founder (with Weldon and Galton) of the journal *Biometrika*. The intention was to provide a vehicle for the mathematical treatment of biological problems. Pearson remained the journal's editor until his death more than 30 years later. In 1911, Pearson, remaining at University College, became its first Galton Professor of Eugenics.

A feature of statistics in the early 20th century was the long-running disagreements between Pearson and his distinguished younger contemporary, Ronald Fisher, in their approaches to data analysis. Pearson favored the use of large data sets, whereas Fisher was prepared to look for causation through small data sets.

As early as 1919, Fisher turned down a job opportunity so as to avoid working with Pearson. However, when Pearson retired in 1933, it was Fisher who succeeded him as head of the Department of Eugenics. In a Solomon-like judgment, Egon Pearson (Karl's son) was simultaneously appointed head of the Department of Statistics.

Pearson was elected a Fellow of the Royal Society in 1896 and a Fellow of the Royal Society of Edinburgh in 1934. He died on April 27, 1936, in Coldharbour, Surrey, England.

—Graham Upton

Further Reading

Porter, T. M. (2004). *Karl Pearson: The scientific life in a statistical age*. Princeton, NJ: Princeton University Press.

Pearson biography: <http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Pearson.html>

PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT

Among his many accomplishments, Sir Francis Galton first introduced the concept of correlation in a book titled *Natural Inheritance*, which was published in 1889. However, Karl Pearson is credited for extending the concept of correlation and for developing the product-moment correlation coefficient. Pearson's product-moment correlation coefficient is by far the most common index of the relationship between two variables, or bivariate relationship.

Pearson's product-moment correlation coefficient measures the degree to which the points in the scatter-plot tend to cluster about a straight line. In other words, the product-moment correlation coefficient *measures the degree of linear relationship between two variables*. If we label the two variables of interest as x and y , then Pearson's product-moment correlation coefficient, denoted by r , is given by the following formula:

$$\begin{aligned} r &= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2}} = \frac{\text{Cov}(xy)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \\ &= \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - n\bar{x}^2} \sqrt{\Sigma y^2 - n\bar{y}^2}}, \end{aligned} \quad (1)$$

where

$$\bar{x} = \frac{\Sigma x}{n} \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n}. \quad (2)$$

The first element of Equation 1 indicates that Pearson r is the ratio of the covariance between variables X and Y to the square root of the product of the x and y variances. Interestingly, only the numerators

of the variance and covariance equation appear in the latter part of Equation 1 because the denominators cancel each other out. Furthermore, Equation 2 could be used to rewrite the Pearson r formula as

$$r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}. \quad (3)$$

However, an even more interpretable way of expressing Pearson r is as follows:

$$r = \frac{\Sigma Z_x Z_y}{n - 1}, \quad (4)$$

where

$$Z_x = \frac{x - \bar{x}}{S_x} \quad \text{and} \quad Z_y = \frac{y - \bar{y}}{S_y}$$

and $S_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \quad \text{and} \quad S_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}.$

That is, S_x and S_y are the standard deviations of the two variables. Thus, Equation 4 indicates that Pearson r is the average cross-product of the standardized x and y variable scores. The fact that Pearson r can be expressed as a product of two variables measured in standard units means that we can express the relationship between two variables even if they are measured on different scales. For example, x can be measured in inches and y can be measured in acres, or in pounds, or in dollars. Provided that both variables are converted to standard units, r can measure their association.

If only the numerator in Equation 4 was used to compute Pearson r , then r would increase as the number of observations in the data set increased, regardless of the true relationship between the two variables. Therefore, $n - 1$ is needed in the denominator of Equation 4 to provide a statistic that is independent of the sample size. This denominator guarantees that r always lies between -1 and $+1$, inclusive.

Pearson's correlation coefficient helps to determine both the magnitude and direction of pairwise variable relationships. The sign of the coefficient tells us whether the relationship is positive or negative.

A positive correlation means that as the values of one variable increases, so do the values of the second variable. Conversely, a negative relationship indicates that an increase in one variable is associated with a decrease in the other variable. The numerical part of the correlation coefficient indicates the magnitude of the relationship. A value of zero indicates no linear relationship between the two variables, whereas the closer the correlation coefficient is to 1 or -1 , the greater the relationship between the variables.

A key assumption pertaining to Pearson r is that the relationship between the two underlying variables is linear. An effective way of assessing the linearity of relationships is via bivariate scatterplots.

Pearson r as a Descriptive Statistic

When Pearson r is needed solely for the purpose of describing a relationship between two variables, then other than the linearity assumption, three other assumptions should be checked before deciding whether the use of Pearson's correlation coefficient is appropriate. These assumptions are as follows:

1. At least one of the variables should represent a continuous variable.
2. Each observation of the dependent variable (Y) must be statistically independent of every other observation.
3. The variability in scores for one variable is approximately the same at all values of the other variable (i.e., the homoscedasticity assumption).

Example With Real Data

Any of the formulae in Equation 1, Equation 3, or Equation 4 can be used to compute Pearson r . Onwuegbuzie was interested in determining a relationship between the total number of points scored (x) by all 30 professional National Football League (NFL) teams and their winning percentages (y) during the 1997–1998 football season. (Both variables are continuous.) These data are presented in Table 1. Also, the scatterplot is displayed in Figure 1.

Table 1 Descriptive Statistics for National Football League Data

<i>NFL Football Team</i>	<i>No. Points Scored (X)</i>	<i>Winning % (y)</i>	x^2	y^2	xy
New York Giants	287	50.00	82369.00	2500.00	14350.00
Washington Redskins	319	37.50	101761.0	1406.25	11962.50
Philadelphia Eagles	161	18.75	25921.00	351.56	3018.75
Dallas Cowboys	381	62.50	145161.0	3906.25	23812.50
Arizona Cardinals	325	56.25	105625.0	3164.06	18281.25
Green Bay Packers	408	68.75	166464.0	4726.56	28050.00
Tampa Bay Buccaneers	314	50.00	98596.00	2500.00	15700.00
Detroit Lions	306	31.25	93636.00	976.56	9562.50
Minnesota Vikings	556	93.75	309136.0	8789.06	52125.00
Chicago Bears	276	25.00	76176.00	625.00	6900.00
San Francisco 49ers	479	75.00	229441.0	5625.00	35925.00
Carolina Panthers	336	25.00	112896.0	625.00	8400.00
Atlanta Falcons	442	87.50	195364.0	7656.25	38675.00
New Orleans Saints	305	37.50	93025.00	1406.25	11437.50
St. Louis Rams	285	25.00	81225.00	625.00	7125.00
New England Patriots	337	56.25	113569.0	3164.06	18956.25
Miami Dolphins	321	62.50	103041.0	3906.25	20062.50
New York Jets	416	75.00	173056.0	5625.00	31200.00
Buffalo Bills	400	62.50	160000.0	3906.25	25000.00
Indianapolis Colts	310	18.75	96100.00	351.56	5812.50
Pittsburgh Steelers	263	43.75	69169.00	1914.06	11506.25
Jacksonville Jaguars	392	68.75	153664.0	4726.56	26950.00
Tennessee Oilers	330	50.00	108900.0	2500.00	16500.00
Cincinnati Bengals	268	18.75	71824.00	351.56	5025.00
Baltimore Ravens	269	37.50	72361.00	1406.25	10087.50
Kansas City Chiefs	327	43.75	106929.0	1914.06	14306.25
Denver Broncos	501	87.50	251001.0	7656.25	43837.50
Seattle Seahawks	372	50.00	138384.0	2500.00	18600.00
Oakland Raiders	288	50.00	82944.00	2500.00	14400.00
San Diego Chargers	241	31.25	58081.00	976.56	7531.25
Totals	Σx 10215.00	ΣY 1500.00	Σx^2 3675819.00	Σy^2 88281.25	Σxy 555100.00
Means	$\bar{X}$ 340.50	$\bar{Y}$ 50.00			

It can be seen from this plot that the relationship is linear, and that homoscedasticity is present. Table 1 presents the summary statistics. From this table, we can use Equation 1 to compute Pearson r , as follows:

$$\begin{aligned}
 r &= \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - n\bar{x}^2}\sqrt{\Sigma y^2 - n\bar{y}^2}} \\
 &= \frac{555100 - 30(340.50)(50.00)}{\sqrt{3675819.00 - (30)(340.50)^2}\sqrt{88281.25 - (30)(50.00)^2}} \\
 &= \frac{555100 - 510750.00}{\sqrt{3675819.00 - 3478207.50}\sqrt{88281 - 75000.00}} \\
 &= \frac{44350.00}{\sqrt{197611.50}\sqrt{13281.00}} \\
 &= \frac{44350.00}{(444.54)(115.24)} \\
 &= \frac{44350.00}{51228.79} \\
 &= 0.87
 \end{aligned}$$

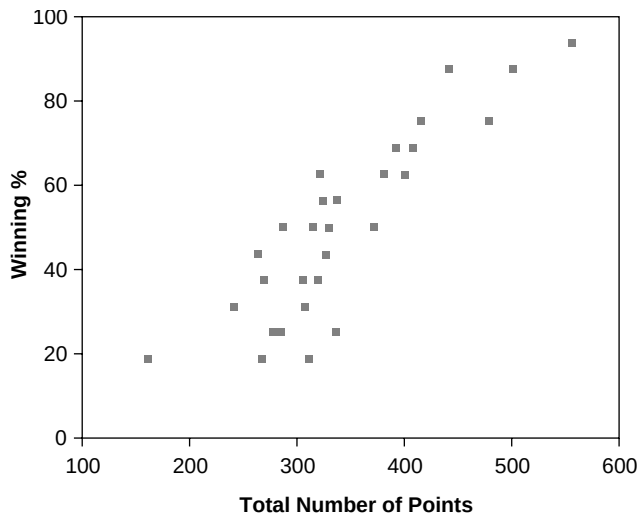


Figure 1 Scatterplot of Number of Points Scored Against Winning Percentage

Thus, $r = .87$ indicates a positive linear relationship between the total number of points scored by NFL teams and their winning percentage, which is consistent with the scatterplot in Figure 1. A useful feature of Pearson r is that if we square it, we obtain r^2 , the coefficient of determination, which indicates that 75.69% (i.e., $.87^2$) of the variance in winning

percentage is explained by the total number of points scored by teams. Although the correlation coefficient is extremely large, we cannot use terms such as *statistically significant relationship* because, as yet, we have not conducted a null hypothesis significance test (NHST) for this correlation coefficient.

Pearson r as an Inferential Statistic

When a researcher's goal is to use the observed Pearson r to make inferences, then a further assumption is needed than for when Pearson r is used for descriptive purposes only—namely, that the dependent variable is normally distributed. The normality assumption should be assessed by both graphical and statistical means. With regard to the former, frequency histograms could be used. In addition to graphical checks of normality, the skewness and kurtosis can be assessed for direction and magnitude. Scores with large positive or negative skewness coefficients and/or large positive (i.e., leptokurtic distribution) or negative (i.e., platykurtic distribution) kurtosis coefficients indicate violations to the normality assumption. It should be noted that large

skewness and kurtosis coefficients affect Type I and Type II error rates.

If the normality assumption holds, the researcher is ready to use Pearson r to determine whether a relationship between two variables is statistically significant—that is, whether the derived Pearson r represents a finding that is sufficiently different from zero to reject the null hypothesis. The NHST concerning a population correlation coefficient (ρ) takes the general form:

$$\rho = \frac{\text{Proportion of variance explained}}{\text{Proportion of variance unexplained}}.$$

Using the fact that R^2 is the proportion of variance explained, and $1 - R^2$ is the proportion of variance unexplained, the test statistic (t) for Pearson r generalizes to

$$t = \sqrt{\frac{R^2}{\frac{1 - R^2}{n - 2}}} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}. \quad (5)$$

The test statistic in Equation 5 is then compared to the associated t critical value based on $n - 2$ degrees of freedom and is determined by the desired level of statistical significance α . The three possible types (two one-tailed and one two-tailed) of NHSTs involving the correlation coefficient and the decision rules for determining whether the null hypothesis is rejected (and a statistically significant relationship inferred) are presented in Table 2.

A close examination of the numerator of the right-hand side of Equation 5 indicates that, holding the correlation constant, as the sample size (n) increases so does the t value, and, consequently, the probability of rejecting the null hypothesis. This probability is

Table 2 Decision Rules for Determining Whether the Null Hypothesis Is Rejected

<i>Alternative Hypothesis</i>	<i>Rejection Region</i>
$H_a: \rho > \text{hypothesized value}$	Reject H_0 if $t > t$ critical value (upper-tail test)
$H_a: \rho < \text{hypothesized value}$	Reject H_0 if $t < -t$ critical value (lower-tail test)
$H_a: \rho \neq \text{hypothesized value}$	Reject H_0 if either $t > t$ critical value or $t < -t$ critical value (two-tailed test)

commonly known as the p value. Conveniently, virtually all statistical software automatically computes this p value, thereby making the use of t tables obsolete. As can be seen from the output generated by SPSS in Figure 2, for the NFL football data set, the p value is less than .0001, which is less than .05 (the typical nominal value). Thus, the null hypothesis that the population correlation is zero is rejected, and we conclude that there is a statistically significant relationship between the total number of points scored by a NFL team and the team's winning percentage.

Effect Size for Pearson r

Because of the influence that the sample size has on p values associated with tests of bivariate correlations, as well as on all other NHSTs, it is not sufficient to

Correlations			
		Winning %	Total Number of Points
Winning %	Pearson Correlation Sig. (2-tailed) N	1.000 . 30	.866** .000 30
Total Number of Points	Pearson Correlation Sig. (2-tailed) N	.866** .000 30	1.000 . 30
**. Correlation is significant at the 0.01 level			

Figure 2 Pearson r for NFL Data

report the r value and p value corresponding to each relationship tested. The magnitude of the r value also should be interpreted. By interpreting the r value, this statistic also serves as an effect size value, resulting in the delineation of the practical significance of the findings. The r value can be interpreted by using Cohen's criteria of .1 for a small correlation, .3 for a moderate correlation, and .5 for a large correlation. Using Cohen's criteria, the relationship found for the NFL data set (i.e., $r = .87$) is extremely large.

—Anthony J. Onwuegbuzie,
Larry Daniel, and Nancy L. Leech

Further Reading

- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r . *The Journal of Experimental Education*, 74(3), 251–266.
- Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). *SPSS for basic statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
- Onwuegbuzie, A. J. (1999). Defense or offense? Which is the better predictor of success for professional football teams? *Perceptual and Motor Skills*, 89, 151–159.
- Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools*, 9(1), 73–90.
- Pearson r Applet: <http://noppa5.pc.helsinki.fi/koe/corr/cor7.html> (allows you to move the cursor on the scale and thereby change the scatterplot)

PERCENTILE AND PERCENTILE RANK

The terms *percentile* and *percentile rank* are considered by some people to mean the same thing. However, these two terms are different in conceptual meaning and should not be used interchangeably.

Percentile Rank

A percentile rank describes the standing, or position, of an earned score in comparison to a reference group. The percentile rank for that earned score indicates the percentage of scores in the reference group that are

lower than the earned score. Thus, an examinee whose test performance earns him or her a percentile rank of 72 would have scored higher than 72 percent of those in the reference group who took the same test.

Percentile ranks are reported in terms of whole numbers between 1 and 99 (inclusive). No decimals are used, and percentile ranks never assume the values of 0 or 100. These extreme values are not used because the percentile rank for any earned score considers that earned score to be *in* the reference group. Considered in that way, it is logically impossible for an earned score to be higher than no scores or all of the scores.

It is important to note that a percentile rank indicates relative (not absolute) standing. Thus, knowing the nature of the reference group is exceedingly important if one is to properly interpret an examinee's percentile. An examinee's percentile rank might be 92 when the reference group is a nationwide group of test takers, 81 when the reference group is defined as those in the examinee's state, but only 47 if the reference group is test takers in the examinee's school.

Percentile

A percentile is a point along the score continuum that divides the reference group's distribution of earned scores into two parts such that a desired percentage of the group's scores lie below that point. For example, if we are dealing with a normal distribution of IQ scores (with a mean of 100 and a standard deviation of 15), then 75% of the IQ scores would fall below 110.1. Thus, that score value (110.1) would be the 75th percentile.

In computing percentiles, we state a percentage (such as 75% in the above example) and then examine the distribution of scores in the reference group to determine what particular value of the score continuum has that particular percentage of scores below it. This sequence of steps is different from that used in determining a percentile rank, for in the latter case, we begin with a given earned score and then determine what percentage of the reference group's scores are below that earned score. Because of this difference,

a percentile can assume a value that's not the same as any earned score. For example, if the earned scores of six examinees on a 20-item quiz in which quiz scores are determined simply as "number correct" are 19, 17, 16, 14, 13, and 12, the 50th percentile would be equal to 15, a score no one earned. (The value of 15 comes from taking the arithmetic average of the two middle scores, 16 and 14.) Pretend that the two lowest-scoring examinees are not part of the group, and in this situation, the 50th percentile for the remaining four scores is 16.5, a score that not only was not earned but also was not earnable!

The primary difference between percentile ranks and percentiles can be conceptualized as a "figure-ground" difference. In both cases, the "ground" is the reference group. The figure, however, is different. With percentile ranks, the figure is an earned score. In contrast, the figure for a percentile is a point on the score continuum.

For any group of scores, there are 99 percentile points. These percentiles divide the scores in the reference group into 100 parts, each containing 1% of the total. Because of this definitional feature of percentiles, certain percentiles are equivalent to other concepts in statistics and measurement. For example, the 50th percentile is equal to the median. Also, the 25th, 50th, and 75th percentiles are equal to the lower, middle, and upper quartile points (Q_1 , Q_2 , and Q_3), respectively. Terciles are the two points that divide the reference group's scores into three equal parts.

Comments

It is important to keep three things in mind when dealing with percentile ranks and percentiles. If these "warnings" are overlooked, it is easy to misinterpret and misuse these concepts.

First, percentile ranks are not the same thing as ranks, even though both involve whole numbers. A rank indicates a person's relative position in a group, with the rank of 1 given to the person who did the best, a 2 given to the person who did second-best, and so on. Thus, a low rank (e.g., 2) indicates that a person has a very high standing in the reference

group. With percentile ranks, high rather than low numbers indicate a high standing.

Second, percentile ranks and percentiles are ordinal in nature and lack the "equal interval" characteristic that's embodied in interval measurement scales. To illustrate this point, suppose four examinees in a larger group of test takers are A, B, C, and D, and further suppose that their percentile ranks on an ability test are 45, 55, 85, and 95, respectively. Because percentile ranks are ordinal, it would be improper to state (or to think) that the difference in ability between Examinees A and B is the same as the ability difference between Examinees C and D. If the reference group's ability scores are normally distributed, the earned test scores from Examinees A and B would be much closer together (along the scale of earnable raw scores) than the earned test scores from Examinees C and D.

Finally, those who interpret test scores need to remember that percentile ranks indicate relative, not absolute, positioning in a group. If an examinee's test performance leads to a percentile rank of 97, for example, that does *not* indicate that he or she correctly answered 97% of the test questions or has very little more to learn. Likewise, a very low percentile rank should *not* be interpreted to mean that an examinee has no knowledge or skill.

Percentiles and percentile ranks are very popular in educational systems. With them, educators can compare different students who take the same test in different backgrounds with their percentiles or percentile ranks in a more reasonable manner. Also, in a large-scale standardized test, students can take past results of the overall data about the relation between raw scores and percentiles or percentile ranks as references to predict their testing achievements.

—Ping-Lun Tu

Further Reading

- Kolen, M. J. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.
- OECD handbook for internationally comparative education statistics: Concepts, standards, definitions and classifications. (2004). Paris: OECD.

PERFORMANCE IQ

Performance IQ is a measure of intelligence that does not require the use of words or language and is associated with the Wechsler Intelligence Scales. When David Wechsler developed his IQ test in 1939, he divided it into two components: tasks that required mainly verbal abilities and tasks that required mainly perceptual-manipulative skills. The Performance IQ can be considered mainly a measure of visual-spatial processing, novel problem solving, attentiveness to visual detail, and visual-motor integration and speed. Recent editions of the Wechsler scales have broken down the Performance IQ into a perceptual organization component and a processing speed component. The perceptual organization scale most closely represents what has been traditionally thought of as Performance IQ.

Another way of thinking about Performance IQ is that it is that aspect of intelligence that does not depend upon experience and learning, sometimes referred to as fluid intelligence. Because experience is not heavily weighted, Performance IQ is considered to be relatively free of cultural bias. For this reason, the Performance IQ is sometimes used in place of the Full Scale IQ with minorities and individuals whose first language is not English as the overall indicator of intelligence. Although the research is far from definitive, there is some indication that skills measured by Performance IQ tend to decline with age, whereas those measured by Verbal IQ tend to stay the same or actually increase. Attempts to tie Performance IQ to the right hemisphere of the brain and Verbal IQ to the left hemisphere generally have not been successful.

—Steve Saladin

See also Wechsler Adult Intelligence Scale

Further Reading

- Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.
- Kaufman, A. S., & Lichtenberger, E. O. (2005). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.

PERFORMANCE-BASED ASSESSMENT

Tests come in many forms, and in a real sense, the act of completing any test items, regardless of their format, constitutes a performance of some type. However, the term *performance-based assessment* (also, *performance assessment* or sometimes *authentic assessment*) has come to possess a specific meaning in the context of measuring cognitive abilities. Although Linn and others have noted that performance tasks are sometimes described relative to what they are *not* (as in, not selected-response, and not norm-referenced), performance-based tests can be characterized fundamentally by extended-length, highly interactive tasks that call for answers or products to be uniquely generated by the test taker. These tasks are often open-ended in nature and, to the extent possible, are framed in contexts approximating real-life applications of knowledge, skills, and abilities. Furthermore, performance-based assessments often involve complex, multistep problems requiring different types of performance, and, depending on the context for testing, may be intended to be completed by individuals working independently or in small groups.

Performance-Based Assessment Methods: Processes and Tasks

Many kinds of activities are commonly considered to be performance-based tasks. A (nonexhaustive) list of performance-based assessment activities includes presentations, research projects, role-plays, Web pages, experiments, portfolios, extended written responses, performance demonstrations, exhibitions, and working through case studies. Although the precise nature of the activity and the format of a performance-based assessment will vary widely by testing program and testing purpose, it is clear that the listed assortment of performance tasks gives test takers the opportunity—within whatever parameters might be defined by the assessment's developer—to produce rather than select. This range can reflect the individuality of test takers and their unique engagement with the specific nature of the assessment task.

Of course, the actual evaluation products could take any of a variety of actual forms, such as something written (here, think of formats such as short or extended-length essays and also reports, brochures, articles, transcripts of conversations, or letters); a live presentation (that may or may not be recorded for multimedia storage and later re-viewing); or the creation of other objects to be assessed according to specific criteria. For example, the creation of a Web site, the production of a working computer program, and the fabrication of a model or prototype object are items to be evaluated that likewise might be generated in response to a performance task. Furthermore, in addition to this range of products that may be generated, some applications of performance-based testing consider the process by which the intended product or answer is created to be on equal footing with the product or answer itself.

Indeed, in some cases, performance-based assessment might be described more accurately as *performance-and-product assessment*. Recognizing that process sometimes is and sometimes is not part of what is evaluated through performance-based assessments, Messick identified three kinds of performance tasks: (a) those in which the performance and product are essentially the same thing (i.e., a dance recital), (b) those where the end product is what mainly counts (i.e., an essay), and (c) those in which the performance and the product are distinctly separate but equally important elements to be evaluated (i.e., a science experiment).

There is one other attribute of performance-based assessment often regarded as a critical aspect of these formats: across applications, performance task users strive to incorporate a high degree of fidelity to realism. Performance tasks tend to be constructed so as to couch questions and activities within situations where the skills, knowledge, and abilities being tested would be commonly used. Think here of the road test for a driver's license: In many states, to make a pass-fail decision about granting the license, examiners ride in cars driven by candidates, on local streets, where other motorists are going about their business. The assessment task in this example provides the scorer with a direct observation of the

underlying skill of interest in a naturalistic setting. Although it could be said that the circumstance of performing for the purpose of being evaluated makes the performance by its very nature artificial, performance-based tasks are intended to inject realism back into the assessment process by setting the performance conditions in a context approximating typical use of the skill to the extent possible.

Although performance-based tasks have long been used by test developers in selected testing contexts such as professional certification and licensure, interest in including performance tasks across test uses has increased over recent years. Indeed, performance-based testing tasks can be appropriately created and adapted for many standardized and classroom testing purposes, from diagnosis and remediation to certification, selection, and accountability. Boodoo and Miller and Linn characterized this broadening use of performance-based assessments as concerted efforts to enhance the depth of measurement information, because performance tasks offer test users the possibility of capturing an array of examinee knowledge and skill in a richer, more realistic context than could be done ordinarily with traditional, selected-response approaches. This is not to say that multiple-choice and other such selected-response formats cannot be used for assessing complex cognitions, but performance tasks are typically characterized by elaborate, multi-step activities that entail considerable investment of time and involvement with the task.

Validity and Reliability Considerations for Performance-Based Assessment

Although substantial realism in measuring the skill or ability of interest and a high level of engagement with the task are among the valued qualities that contribute to making performance-based assessments appealing to many, there are likewise several psychometric issues that users should keep in mind with respect to the development, scoring, and use of such tasks. Performance-based assessments are unique enough that there are specific considerations that users need to take into account when they implement such tests

for most large-scale testing purposes, and many classroom applications as well.

Validity concerns specific to performance-based assessment have been the focus of many researchers. Although integrating a high degree of realism is an important part of performance-based assessment, the appearance of fidelity to real-life contexts (i.e., face validity) is not enough. The expectation is that performance tasks have meaning, require complex cognitions applicable to real-life problems, minimize the extent to which they draw on ancillary skills, and are rated according to explicit standards. Linn, Baker, and Dunbar describe several potential sources of validity-related evidence supporting the proposed uses of performance tasks, including consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency.

Another measurement issue for performance-based assessment concerns generalizability across tasks and topics and task specificity. As mentioned above, administrative limitations on testing time often result in performance-based tests that consist of just one or a very small number of tasks. This, in turn, offers a limited sample of behaviors to generalize to the larger domain of knowledge being tested, and it calls into question how broadly or narrowly the larger domain is defined and the extent to which the submitted performance is representative of (or generalizes to) tasks in that larger domain. Similarly, the limited number of tasks included in many performance assessments raises the potential for examinee by task interactions: Because examinees interact with most tasks in some depth for an extended length of time, care must be taken to ensure that tasks are general enough so that the performances are not affected negatively by an unfairly constrained context. For example, on a writing assessment where the skill of interest is descriptive writing, rather than insisting that all test takers expound about highly individual interest-specific topics such as ice hockey or favorite Montana resort towns, tasks should be defined more broadly to ensure that each test taker can reasonably draw from personal experience to complete the task (such as writing about hobbies enjoyed or vacations taken). These concerns

of generalizability and task specificity can be viewed as potential sources of invalidity because they can limit the extent to which inferences about individuals' proficiency in particular content areas can be made.

Another important consideration for performance-based assessment is reliability. Reliability concerns the consistency of test scores across both the administration of testing tasks and the scoring of test takers' responses. In terms of reliability, many performance-based tests are often composed of just one or a small number of highly involved tasks, and although examinees may engage in those tasks in great depth for an extended length of time, in terms of actual measurement opportunities, the yield may be minimal. For this reason, performance-based assessments may not appear to measure up favorably to other assessment formats in terms of traditional reliability expectations, because statistical approaches to reliability (such as computing internal consistency and split-halves) try to estimate how much variability in scores is due to random and systematic error, and how much is due to true ability. These indices benefit from lots of observations: When low reliability estimates are observed on tests composed of multiple-choice items or other such selected-response formats, one common solution is to increase the number of items administered. But because performance tasks are often elaborate in terms of both time and resources, adding tasks may not be feasible in some performance-based testing contexts for practical reasons. In this way, the validity challenge for test developers is to balance traditional expectations about efficiency, reliability, and comparability with other desired assessment qualities advanced by the use of performance-based assessments (e.g., fidelity, complexity, and response freedom).

Whether the performances submitted for evaluation are to be scored using holistic or analytical scoring rubrics, test developers must work to standardize the scoring process as much as possible to ensure that scorers apply the standards for judging performance in a consistent way, and to reduce the opportunities where subjectivity may enter into the scoring process. Procedures should be in place to train individual scorers and thoroughly familiarize them with the rubric being used (including what it does and does not

include), focusing on how to apply it uniformly across all of the performances they rate. Scorers should be monitored throughout for systematic and random deviations from the rubric, and spiraling in previously scored anchor responses can be helpful in this regard for monitoring purposes. Where necessary, remediation should be provided to realign scorers with the rubric. It should be explicit in the rubric what factors of the performance are to be reflected in the score and which are defined as extraneous. For example, is spelling or neatness explicitly defined as a part of the construct being assessed and therefore given weight in scoring? Cost permitting, each performance should be considered by more than one scorer, and another aspect of consistency for test developers to monitor is the extent to which scorers provide ratings of the performances that are consistent with one another. By putting these kinds of measures in place, comparability in scoring across different test takers' performances and across scorers can be maximized to the extent possible.

Conclusions

Ultimately, when performance-based assessments are constructed carefully, administered appropriately, and interpreted properly, they can provide a distinctive, in-depth window into how individuals apply knowledge, skills, and abilities in realistic ways. The possibilities for what can be evaluated encompass a wide range of products and processes, positioning performance-based assessment at the least-constrained end of the continuum of assessment methods ranging from most constrained (multiple-choice and other selected-response formats) to least constrained. Of course, as assessments, performance tasks must aspire to high standards of test quality, but by balancing methodological rigor with construct-driven task development, performance-based assessments permit test takers to engage more thoroughly in many respects with assessment tasks than most tests generally permit, and thus offer test users perspectives on proficiency informed by such engagement, to the benefit of test takers and users alike.

—April L. Zenisky

Further Reading

- Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, 48(12), 1210–1218.
- Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (RR-90-7). Princeton, NJ: Educational Testing Service.
- Boodoo, G. (1993). Performance-based assessment or multiple choice? *Educational Horizons*, 72(1), 50–56.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance-based assessments. *Applied Measurement in Education*, 4(4), 289–303.
- Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In E. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237–270). Washington, DC: American Council on Education.
- Linn, R. L. (1995). High stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In T. Oaklund & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 49–74). Boston: Kluwer.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, 13(1), 5–8, 15.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Miller, M. D., & Linn, R. L. (2000). Validation of performance assessments. *Applied Psychological Measurement*, 24(4), 367–378.

PERITZ PROCEDURE

The Peritz procedure can be applied after a significant overall F test in a one-way ANOVA. A series of additional F tests can be used to determine the significance of the difference between each of the possible pairs of means. Among k means, there will be $k(k - 1)/2$ pairs of means. If the usual assumptions of an ANOVA are satisfied, then the probability of one or more Type I errors is limited to the nominal level, α , of the overall F test. This will be true when each mean is based on the same sample size, N , and when sample sizes differ no matter how large the difference

in sample sizes may be. For a wide variety of conditions, the Peritz procedure is more powerful than any other procedure for pairwise testing of means.

The application of the Peritz procedure can be accomplished by testing all partitions of the k means into subsets. For a particular pair of means to be found significantly different, it must be the case that a significant F test is found for every partition of the k means in which the two means in question are included in the same subset. However, progressively smaller subsets must be tested at more stringent levels. For example, if five means are partitioned into two subsets of the first three means and the last two means, a separate F test would be applied to the two subsets. The first subset would be tested at level $(3/5)\alpha$ and the second subset at level $(2/5)\alpha$. That is, the first subset of three means includes three of the five means. It is tested at $3/5$ times α . The second subset of two means includes two of the five means. It is tested at $2/5$ times α .

If either of these tests is significant at the designated level, then the partition is rejected. Nonsignificance of the partition requires that a number of pairs cannot be significantly different. For example, there are three pairs among the means in the first subset (1, 2), (1, 3), and (2, 3). There is also the single pair in the second subset, (4, 5). All four pairs must fail to differ significantly if both F tests for this partition fail to be significant. Of course, all other partitions must be tested as well.

If the five means are partitioned into three subsets of the first two means, the third and fourth means, and the single fifth mean, then the first two means would be tested at level $(2/4)\alpha$ and the second subset at $(2/4)\alpha$. The denominator of each fraction is taken as the number of means included in subsets of at least two means.

Slightly greater power can be obtained for the Peritz procedure by using slightly less stringent levels in each subset of the partition. For example, consider the case of five means partitioned into two subsets of the first three means and the last two means. Instead of testing the first subset of three means at level $(3/5)\alpha$, it is possible to test that subset at level $1 - (1 - \alpha)^{3/5}$. Similarly, instead of testing the second subset at level $(2/5)\alpha$, it would be tested at level $1 - (1 - \alpha)^{2/5}$. Again, the

Table 1 Data for Five Groups in an ANOVA

	Control	Treat 1	Treat 2	Treat 3	Treat 4
Mean	4.05	3.52	3.15	3.18	2.24
N	42	24	38	40	40

partition would be rejected if and only if at least one of the F tests is significant at the designated level.

Consider the data in Table 1, in which four treatment groups are being compared to each other and to a control. Lower scores indicate better performance.

The MS_{WG} for these data is 3.0326 with $df_{WG} = 179$. The ANOVA for these data would give $F = 5.75 > 2.42 = F_{.95}(4, 179) = CV$. Therefore, we reject the full null hypothesis at the .05 level and proceed to pairwise testing. Suppose we consider the requirements for finding means 3.18 and 2.24 to be significantly different.

A total of 15 possible partitions must be tested and found significant in order for the difference between these two means to be significant. The first partition is the one in which all five means are together and the full null is tested. In the present case, the second partition has means 1, 2, and 3 together, and means 4 and 5 form a second subset. As shown in Table 2, the second subset of two means also produces a significant F at the appropriate level.

The differences between ordered means are in Table 3. The largest difference is 1.81 between Treatment 4 and the control. The individuals in Treatment 4 have significantly lower average scores than does the control. The second largest difference is 1.28. Those in Treatment 4 are significantly lower than those in Treatment 1.

The difference of .94 between Treatment 4 and Treatment 3 is significant, as demonstrated by the 15 partitions shown above. The original Newman-Keuls procedure cannot be applied because it does not provide accurate control of Type I errors. This difference of .94 would not be found significant when the Shaffer-Welsh modification is applied to the Newman-Keuls using the harmonic mean of sample sizes. The Peritz procedure is shown to be more

Table 2 Partitions for Testing Groups 4 and 5 of Table 1

Partition	Grouping	Group Size	F	CV
1	1 1 1 1 1	5	6.42*	2.422
2	1 1 1 2 2	3	2.70	3.566
		2	5.83*	5.483
3	1 1 2 1 1	4	7.65*	2.655
4	1 1 2 2 2	2	1.41	5.483
		3	3.74	3.566
5	1 1 2 3 3	2	1.42	5.087
		2	5.83*	5.187
6	1 2 1 1 1	4	7.38*	2.655
7	1 2 1 2 2	2	5.33	5.483
		3	4.91*	3.566
8	1 2 1 3 3	2	5.33*	5.087
		2	5.83*	5.087
9	1 2 2 1 1	3	11.07*	3.566
		2	0.66	5.483
10	1 2 2 2 2	4	3.47*	2.655
11	1 2 2 3 3	2	0.66	5.087
		2	5.83*	5.087
12	1 2 3 1 1	3	11.07*	3.046
13	1 2 3 2 2	3	4.91*	3.046
14	1 2 3 3 3	3	3.74*	3.046
15	1 2 3 4 4	2	5.83*	3.894

* $p < .05$.**Table 3** Differences Between Ordered Means of Table 1

	Tr 4 2.24	Tr 2 3.15	Tr 3 3.18	Tr 1 3.52	Control 4.05
Tr 4 = 2.24	—	.91	.94*	1.28*	1.81*
Tr 2 = 3.15		—	.03	.37	.90
Tr 3 = 3.18			—	.34	.87
Tr 1 = 3.52				—	.53

* $p < .05$.

powerful while keeping the Type I error control to be exact even with unequal sample sizes.

—Philip H. Ramsey

Further Reading

Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods*, 7, 504–523.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range test. *Journal of Educational Statistics*, 4, 14–23.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, 72, 566–575.

PERSONAL PROJECTS ANALYSIS

Personal projects are extended sets of personally relevant action with important implications for human well-being and adaptation. Personal projects analysis (PPA) is a methodology that allows researchers and practitioners to assess the content, appraisal, hierarchical nature, and impact of personal projects. Developed by Brian R. Little in the late 1970s, PPA has been widely adopted as the standard for assessing personal action and goal constructs in diverse fields of study. It is not a test, but an integrated set of flexible assessment modules. PPA is based upon a set of 12 foundational criteria that are prescribed as important to meet before orthodox psychometric criteria such as reliability and validity are addressed. Eight of these are discussed here as being particularly relevant to quantitative assessment methodology.

By way of brief overview, PPA provides a set of modules for (a) eliciting an individual's personal projects, (b) appraising these projects on a set of dimensions that has been selected for both theoretical relevance and applied utility, and (c) assessing additional features of project systems such as their hierarchical structure and impact. We will first discuss the foundational assumptions of the methodology, then describe two of the modules, and conclude with brief comments on the significance of the methodology for contemporary basic and applied research.

Eight Foundational Criteria for PPA

Personal Saliency

The personal saliency criterion requires that the units be personally significant to the individual being assessed and that these units be expressed in their own idiosyncratic language. In contrast with standard

inventories, PPA does not supply instances of the unit of analysis, but elicits them. The personal saliency criterion ensures that individuals can generate examples of the plans and actions they deem important to know if we are to understand their lives.

Ecological Representativeness

Projects are not just personal constructions, they are contextually embedded actions. This criterion requires that an assessment procedure afford an understanding of the person's everyday social ecology. Whereas the personal saliency criterion highlights the subjective importance of personal projects, this criterion illuminates the contextual factors that frustrate or facilitate project pursuit.

Social Indicator Potential

Besides providing information relevant to individuals, we believe that assessment can also provide information relevant to the social ecologies (e.g., workplace, rural area) within which individuals carry out their projects. In short, data derived from PPA have the potential to contribute to the development of social indicators—such as which projects are difficult for elderly people living in rural areas, or which are most enjoyable for inner-city youth. This criterion ensures that the data we gather on individuals and their pursuits are relevant to policy makers who design the contexts within which these pursuits take place.

Temporal Extension

Many assessment devices use units that take “snapshots” of human characteristics; PPA allows the temporally extended nature of daily lives to come into focus. Personal projects typically proceed through stages of inception, planning, action, and termination, although there are individual differences in the extent to which these stages are negotiated. Although much of the research on personal projects is cross-sectional, there is increasing use of longitudinal designs using hierarchical analysis in which we can examine projects embedded within persons, who in turn are embedded in different ecologies.

Joint Idiographic and Normative Measurement

PPA was based on the assumption that both individual-level and aggregate-level information are important aspects of measurement. When projects are rated on appraisal dimensions such as stress and enjoyment, we can measure the relationship between these variables within the single case (measuring across the individual's projects) or between cases (using the mean of the ratings on each dimension per individual). Unnecessarily divisive debates have characterized assessment methodology in terms of whether the appropriate measurement focus ought to be on idiographic or normative concerns. PPA assumes that each approach provides distinctive information of value to the basic researcher and applied analyst.

Systemic Measurement

In their daily lives, people are typically engaged in multiple projects, and these projects form systems. This criterion requires that the measurement device provide ways in which these systemic features of our unit of analysis can be accessed. PPA modules allow us to explore the conflicts and synergies that characterize complex human lives. Although projects are the equivalent of test items in conventional measurement, it is not meaningful to consider the systemic relations among test items in the same way as one does with projects. The latter have a reality status that is discernibly different from that of items.

Modularity

The modularity criterion requires a method to be flexible enough that it can be adapted readily to research situations. For example, unlike conventional testing, where items and scales are fixed, the dimensions upon which individuals appraise their projects can be selected from a standard set of approximately 20 that have broad theoretical utility, or create ad hoc dimensions specially chosen to reflect the particular person and social ecological context being assessed. The appraisal matrix therefore serves as a kind of

system board into which standard or special dimensions can be plugged for appropriate use. Among ad hoc dimensions used in PPA have been “age you feel when engaged in this project” in a study of mid-life transitions and “extent of benefit to others” in studies of volunteering.

Integrative Units

Assessment of cognitive, affective, and conative aspects of human conduct is typically carried out with different instruments, each with its own unit of analysis. In contrast, personal projects provide a natural base for integrating these interrelated features of daily life. In this respect, we think of personal projects as carrier units—a given project carries information about construed action, experienced affect, and challenges to volitional capacity. Similarly, projects carry information about both the individual and the context and can be regarded as transactional units of analysis. The integrative criterion is central to the project analytic endeavor because one of the central problems in contemporary social research is the difficulty of conjoining information derived from different specialty areas. By its modular, systemic, and integrative features, personal projects analysis offers direction for synthesis in diverse fields.

Assessment Modules in PPA

There are four major assessment modules used in PPA, two of which are particularly relevant to measurement issues—the project elicitation and project appraisal modules.

Project Elicitation Module

This requires respondents to generate a list of their personal projects. These can range from “taking out the trash” to “trashing a political opponent.” The content of projects listed is informative and can be examined through traditional procedures of content analysis. The most frequently elicited personal projects are interpersonal projects and either academic or work projects.

Project Appraisal Matrix

In this module, the participant rates each personal project on a set of standard as well as any desired ad hoc dimensions.

More formally, a $j \times k$ matrix is completed by the participant, in which j personal projects are rated on an 11-point Likert scale that ranges from 0 to 10, across k features or dimensions (such as Importance, Difficulty, Challenge, Stress, etc.). Both j and k are variable, with j typically being between 8 and 12, and k between 17 and 24, although these constraints can be varied to suit the particular research question. The data can be analyzed at the single-case level or normatively, the latter approach involving the derivation of k total scores (often taken as the mean of each of the k features across the j projects).

PPA dimension mean scores are often used as predictor variables with diverse measures of well-being, flourishing, and competency as outcome measures. Five theoretically relevant factors emerged from early factor-analytic studies of the correlations among project dimensions: project meaning, structure, community, efficacy, and stress. In meta-analyses of studies using these factors as predictors of well-being and depression, project efficacy and (absence of) project stress were the most reliable predictors. In recent years, in response to the interest in affective processes and project pursuit, we have augmented the appraisal matrix with a larger number of positive and negative affect dimensions. Factor analyses of these matrices have reliably produced dimensions of project meaning, manageability, support, positive affect, and negative affect.

The use of PPA at both idiographic and normative levels raises some important psychometric issues. Until the late 1990s, the studies of PPA that used factor analysis had primarily examined mean-level data within the same study, and so the extent to which between-study differences at the project-system level exist had not been mapped out. Some project-level data had been so analyzed, but results were viewed as possibly biased because of the violation of the assumption of independence of observations. Little pointed out the possibilities inherent in the ipsative

analysis of PPA features and noted that individual correlations within a matrix do not necessarily follow the pattern that obtains normatively. This problem is known more generally as Simpson's paradox, wherein generalizations made on the basis of data analyzed at one level of aggregation may be invalid at different levels of aggregation. For continuous data, correlations computed at the normative level do not logically imply that such correlations must obtain at the individual level. The magnitude, and even the sign, of the correlations can theoretically change when the data are disaggregated, and the mapping is an empirical, rather than a logical, issue. Travis Gee, in a detailed psychometric analysis of the properties of what he termed "project space," demonstrated that there is, in fact, a high degree of isomorphism between the relationship among dimensions analyzed normatively and those analyzed at the individual level. Indeed, in the rare cases where there is a lack of fit with the normative project space, one might be detecting a case of clinical interest.

Implications for Contemporary Basic and Applied Research and Practice

PPA has generated a number of related techniques for the assessment of life tasks, personal strivings, and personal goal pursuit. These have been generally referred to as personal action construct (PAC) units and are a legitimate alternative to more conventional units of analysis, such as traits. PAC units are, in some respects, closely related to some of the social cognitive learning units of analysis that are flourishing both as theoretical constructs and as aids to clinical assessment. Personal projects and other PAC units, however, are pitched at a somewhat higher level of molarity. A priority for both theoretical and applied research is the integration of middle-level units and more molecular cognitive social learning units.

At the applied level, researchers have adopted or adapted PPA to explore problems in areas as diverse as environmental planning, developmental transitions, and moral philosophy. Among the most active adopters of PPA methodology have been researchers

and practitioners in rehabilitation and occupational therapy, where personal projects have been adopted to explore the factors that lead to therapeutic compliance and coping with illness. Notable too, at the other end of the spectrum of human well-being, is that human flourishing can be understood as the sustainable pursuit of core projects. In short, personal projects analysis is a flexible suite of methodological tools designed for the integrative study of human lives in the worst of times and the best.

—Brian R. Little and Travis L. Gee

Further Reading

- Gee, T. (1998). *Individual and joint-level properties of personal project matrices: An exploration of the nature of project spaces*. Unpublished doctoral dissertation, Carleton University, Ottawa, Canada.
- Little, B. R. (1983). Personal projects analysis: A rationale and method for investigation. *Environment and Behavior*, 15, 273–309.
- Little, B. R. (1989). Trivial pursuits, magnificent obsessions and the search for coherence. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions*. New York: Springer-Verlag.
- Little, B. R. (2005). Personality science and personal projects: Six impossible things before breakfast. *Journal of Research in Personality*, 39, 4–21.
- Little, B. R., Salmela-Aro, K., & Phillips, S. D. (Eds.). (in press). *Personal project pursuit: Goals, action and human flourishing*. Mahwah, NJ: Erlbaum.

Personal Project Analysis, by Brian R. Little: <http://www.brianrlittle.com/ppa/index.htm>

PERSONALITY ASSESSMENT INVENTORY

The Personality Assessment Inventory (published by Psychology Assessment Resources) is a self-administered, objective measure of adult personality and psychopathology designed to aid in the identification of critical client variables in therapeutic settings. The PAI consists of 344 items and four sets of scales: four validity scales (Inconsistency, Infrequency, Negative Impression, and Positive Impression);

11 clinical scales covering the major categories of pathology corresponding to the *DSM* nosology (i.e., Somatic Complaints, Anxiety, Anxiety-Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, and Drug Problems); five treatment scales measuring constructs related to treatment and case management (i.e., Aggression, Suicidal Ideation, Stress, Nonsupport, and Treatment Rejection); and two interpersonal scales (i.e., Dominance and Warmth). Respondents are asked to read each item and decide the self-accuracy of the statements on a 4-point Likert-type scale (i.e., False–Not at All True, Slightly True, Mainly True, or Very True). The PAI generally takes respondents 45–50 minutes to complete.

Unlike similar objective measures of personality and psychopathology (e.g., Minnesota Multiphasic Personality Inventory-2), all of the PAI scales are independent, with no item overlap. Among the combined normative sample, test-retest reliability of PAI clinical scales over an average of 24 days ranged from .79 to .92. This reflects the fact that the instrument taps relatively enduring patient characteristics rather than current clinical state alone. Validity studies have generally found the PAI scales to represent the domains of interest, sufficiently discriminate among clinical symptoms, and correlate with alternate existent measures of personality and psychopathology. Comparative norms exist for general adults, adult psychiatric inpatients, and college students (ages 18 and older).

Although a relatively new assessment instrument, the PAI has gained wide support from researchers, professional trainers, and practitioners. Since its introduction in 1991, the PAI has been included in more than 300 published studies. In addition, the PAI ranks fourth among personality tests employed by graduate training programs and predoctoral internships accredited by the American Psychological Association. With regard to practice applications, and in addition to traditional psychotherapeutic settings, the PAI is widely used in forensic settings and in legal cases involving emotional injury.

—Jacob J. Levy

Further Reading

- Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychology Assessment Resources.
- Morey, L. C. (1996). *An interpretive guide to the Personality Assessment Inventory*. Odessa, FL: Psychology Assessment Resources.
- Morey, L. C. (2003). *Essentials of PAI assessment*. Hoboken, NJ: Wiley.

Psychology Assessment Resources: <http://www.parinc.com>

PERSONALITY RESEARCH FORM

The Personality Research Form (published by Sigma Assessment Systems) was developed by Douglas N. Jackson to assess Murray's comprehensive system of human needs. The PRF can be viewed as tapping fundamental features of motivation with broad relevance to human behavior across diverse contexts.

Various forms of the PRF are available. A and B are parallel forms and shorter, with only 15 scales. Forms AA and BB are also parallel forms and, like E, contain an additional 7 scales. Form G is the same as Form E but does not contain the Infrequency scale. The scales for forms A, B, AA, and BB contain 20 items each, whereas the scales of E and G contain 16 items each. Forms E and G are by far the most commonly used. The PRF-E contains 20 content scales for assessing human needs and two validity scales (Infrequency and Desirability). Infrequency was designed to assess random or careless responding or invalidity due to poor language comprehension. The Infrequency items are unlikely but not impossible. A score of 4 out of 16 indicates invalidity. Desirability is a measure of the evaluative good-bad dimension in which low scores can indicate an insensitivity to evaluation and high scores represent a propensity to appear favorable. The Desirability scale was also used for item selection. There is no item overlap between any of the 22 scales, and, with the exception of Infrequency, items were selected that correlated more highly with their own scale than any other scale.

The 22 scales of Form E contain 8 items true keyed and 8 items false keyed. Items are ordered to foster local independence by presenting one item from each of the 22 scales, and then this block of 22 items repeats in the same sequence with alternative keying. The 20 bipolar content scales of Form E can be categorized into the following: (a) impulse expression and control (Impulsivity and Change vs. Harm Avoidance), (b) work and play (Achievement and Endurance vs. Play), (c) direction from other people (Succorance vs. Autonomy), (d) intellectual and aesthetic pursuits (Understanding and Sentience), (e) ascendance (Dominance vs. Abasement), and (f) interpersonal relationships (Affiliation, Nurturance, Exhibition, and Social Recognition vs. Aggression and Defence).

The PRF can be administered in paper-and-pencil format with reusable question booklets to individuals or to groups and takes about 35 minutes. In addition to hand scoring, mail-in scoring, and Internet scoring, computerized administration with automated scoring and reporting is also available. Software reports are either clinical with some interpretation or ASCII (text) data file reports for research. The PRF is available in English, French, and Spanish. The reading level is between Grades 5 and 6. Normative data stratified by gender are available separately for juveniles and adults.

—John R. Reddon

See also Jackson, Douglas N.; Personality Assessment Inventory

Further Reading

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, 78, 229–248.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Personality Research Form: <http://www.sigmaassessmentsystems.com/assessments/prf.asp>

features of one's character that influence behavior across different situations and settings. In measuring these internal and manifested features, personality tests are used to assess one's affective, motivational, interpersonal, and attitudinal characteristics, as opposed to one's abilities or achievements. Personality tests can be categorized by the method in which data are obtained (e.g., self-report vs. projective methods) as well as by the type of personality information one wishes to assess (e.g., psychopathological personality assessment vs. nonpathological or normal personality assessment). This entry will provide a brief overview of the issues pertinent to both methods and types of personality tests.

Self-Report Personality Tests

Personality is most commonly assessed by asking respondents to endorse, rank, or otherwise acknowledge that they possess certain characteristics, hold certain attitudes and beliefs, or exhibit certain behaviors by administering one or more self-report personality inventories. Self-report measures provide specific stimulus materials, usually in terms of questions, statements, and descriptor lists. For example, one of the earliest self-report personality tests is the Adjective Checklist. This test consists of a list of more than 300 adjectives or personality descriptors in which the respondent is asked to "check" the ones that are generally accurate descriptions of themselves (examples of adjectives are *happy*, *sad*, *anxious*, *depressed*, *friendly*, etc.). Over the years, self-report inventories expanded beyond simple descriptors to include questions and statements about beliefs, attitudes, and behaviors, including broad measures of personality (e.g., the Minnesota Multiphasic Personality Inventory, the 16 Personality Factors Questionnaire, California Personality Inventory, etc.) as well as more narrowly focused personality measures (e.g., Beck Depression Inventory, State-Trait Anxiety Inventory, State-Trait Anger Expression Inventory).

Development of self-report measures generally involves some combination of rational (or theoretical) empiricism. Items for self-report tests are generated

PERSONALITY TESTS

The term *personality* is conventionally defined in broad terms to reflect the enduring internal and manifested

because they fit what one rationally (or theoretically) believes represents the domain of interest. For example, if one is interested in developing a self-report measure of extraversion, one would begin by constructing items that rationally describe what it means to be extraverted (e.g., I enjoy interacting with a large group of people; I am a very sociable person). Other instruments are designed to measure a specific theory of personality. For example, items on the Millon Clinical Multiaxial Inventory reflect Theodore Millon's theory of personality. Once items are conceptualized (rationally or theoretically), they are then empirically examined within the population. Items are included on the final version of the test because they discriminate consistently and significantly between people who possess or represent the domain of interest (e.g., extraverts) and those who do not possess or represent the domain of interest (e.g., introverts). Thus, self-report personality tests rely heavily on psychometric properties (i.e., reliability, content and construct validity).

One of the most critical issues regarding self-report personality testing is unfair test bias. Test bias is defined as the statistical difference found on test results among groups (e.g., difference found between women and men, difference found among racial or ethnic groups). Bias becomes unfair when the differences in test results among groups are not due to actual difference in the domain of interest, but rather due to some extraneous factor. For example, if the domain of interest is competitiveness, and the test item states, "I like to race in regattas," only people who have knowledge of and had opportunities to engage in yacht racing would endorse this item. Respondents without such knowledge of and opportunities to participate in regattas would not endorse this item, but very well may be as competitive as those who do endorse this item. In this case, the bias may be unfair on the basis of socioeconomic status, and possibly race, ethnicity, gender, and geographic origin. That is not to say that the item does not measure the domain of interest, but rather that the item has differential validity based on various social, cultural, and demographic variables. Thus, when selecting a self-report personality test, it is crucial to inquire about the standardization sample on which the

test was normed, as well as which populations the test has been found to appropriately represent. In addition, self-report inventories are not always appropriate for groups of people who may have difficulty understanding or responding to test material (e.g., children, the elderly, or people with below-average cognitive functioning). In these cases, alternative means of assessment may include interviews, behavioral assessments, or other-report inventories (from teacher, parents, partner, peer).

Projective Techniques

Another method of assessing personality avoids the potential pitfall of having test material that may be more or less familiar to certain groups—projective techniques. Projective personality tests involve presenting respondents with ambiguous stimuli (e.g., inkblots, pictures) and having them provide (or project) meaning or interpretation for the stimuli. The underlying assumption of this method is that the way in which the respondent perceives and interprets the ambiguous stimuli reflects one's personality. Thus, the respondent is free to project any thoughts, feelings, attitudes, or beliefs onto the stimuli and is not limited to responding to a specific set of questions or statements.

Projective techniques evolved out of the psychoanalytic tradition. One of the defense mechanisms for anxiety identified by Sigmund Freud was projection, which is described as placing (or projecting) anxiety-arousing impulses onto other objects (in Freud's theory, these objects were other people). For example, if one is frightened by seeing a snake while walking in the woods with a friend, one may react by saying, "We need to leave because you are scared." Projective techniques operate in a similar manner. For example, on the Thematic Apperception Test, pictures of people in ambiguous situations are presented, and respondents are to tell a story about what is happening in the picture. It is believed that the respondents will project their own feelings, attitudes, and beliefs onto the characters in the pictures.

Although many projective techniques are not scored numerically, as with self-report measures, psychometric

properties remain relevant. Interpretations of one's responses to projective materials need to be consistent across test administrators (i.e., interrater reliability is crucial). With regard to validity, projective stimuli rarely (if ever) have content validity because they are designed to be ambiguous. The type of validity that is most crucial in projective testing is criterion-related validity (i.e., concurrent and predictive). Results of projective testing should yield correspondence to other evidence associated with the domain of interest and should also be predictive of future behavior. Potential test bias is also a critical issue; however, the nature of test bias in projective testing differs from self-report inventories. For projective techniques, bias does not exist with the item or stimulus object, but rather in the interpretation of responses. In interpreting responses to projective material, consideration of cultural, social, and other salient factors must be included. The lack of agreement among professionals about how to interpret projective techniques remains one of the biggest problems with this method of personality testing.

Testing for Psychopathology

In clinical contexts, personality testing is typically used to assess for the presence of psychopathology as well as issues of emotional adjustment. The most widely used measure of personality and psychopathology is the Minnesota Multiphasic Personality Inventory (MMPI; now in its second edition—MMPI-2; also available in an adolescent version—MMPI-A). This is a self-report test designed to assess a broad array of psychological symptoms and is primarily employed as a diagnostic aid (it should be noted that no single test or assessment technique should be the only criterion for assigning psychiatric or psychological diagnoses). Other commonly used broad self-report measures of personality and psychopathology are the Personality Assessment Inventory (PAI); the Personality Diagnostic Questionnaire (PDQ; now in its fourth edition—PDQ-4); the Personality Inventory for Children (PIC; now in its second edition—PIC-2; this inventory is completed by the child's parent or guardian); and the Millon Clinical Multiaxial Inventory

(MCMI, now in its third edition—MCMI-III; this measure has a strong focus on assessing personality disorders). In addition to broad measures of personal pathology and psychopathology, many self-report measures have been developed to assess for specific psychopathological symptoms and disorders. Some of the most common self-report measures used to assess narrow personality constructs are the Beck Depression Inventory (BDI; now in its second edition—BDI-II); the State-Trait Anxiety Inventory (STAI); the State-Trait Anger Expression Inventory (STAXI; now in its second edition—STAXI-2); and the Hamilton Depression Rating Scale (HAMD).

Several projective techniques are also commonly used to assess personality and psychopathology. Projective techniques may be preferable to self-report measures with clients who may have a tendency to misrepresent themselves (regardless of intention). Because projective techniques lack face validity, it is difficult for respondents to fake the "correct" answer. Also, the content of projective responses is generally less important than the themes or patterns of responses. The two most popular projective techniques used to assess for psychopathology are the Rorschach Inkblot Method (RIM) and the Thematic Apperception Test (TAT). Other common projective techniques include sentence completion tests, graphic techniques, and assessments of early memories.

Nonpathological Personality Testing

Personality tests are not always used to assess psychopathology or what is "wrong" with one's personality. Many tests have been developed to assess one's "normal" personality or how one generally interacts with his or her environment. This type of nonpathological, or normal, personality testing arose out of the work of trait theorists, such as Gordon Allport, Raymond Cattell, and others. The underlying philosophy of trait theory is that all people are born with distinguishing personal characteristics and qualities called traits. These traits interact with one's environment, resulting in a variety of outcomes (e.g., satisfying or unsatisfying). The current state of the field generally endorses a five-factor model of personality

(referred to as the “Big Five”): Neuroticism (also referred to by its inverse, Emotional Stability); Extraversion; Openness; Agreeableness; and Conscientiousness. Some of the most popular tests for assessing one’s normal personality traits include the NEO Personality Inventory (NEO-PI, now revised—NEO-PI-R), the 16 Personality Factor Questionnaire (16PF), and the Myers-Briggs Type Indicator (MBTI). In addition to trait-based tests of normal personality, other tests have also been designed to measure salient psychological issues in normal populations. The most prominent of these measures is the California Personality Inventory (CPI). One context in which these types of personality tests have been very popular is in educational and vocational counseling.

—Jacob J. Levy

Further Reading

- Anastasi, A., & Urbina, S. (1997). Part four: Personality testing. In *Psychological testing* (8th ed., pp. 348–471). Upper Saddle River, NJ: Prentice Hall.
- Hilsenroth, M. J., & Segal, D. L. (Eds.). (2004). *Comprehensive handbook of psychological assessment (Volume 2): Personality assessment*. Hoboken, NJ: Wiley.

PIE CHART

Pie charts are used to illustrate proportions for categorical variables (nominal or ordinal data). An entire pie represents 100% of the measured variable. The sectors of a pie are proportional to the total number. Negative numbers and percentages over 100 are not presented in a pie chart. These charts are sometimes labeled as 100%, cake, circle, or percentage charts.

Figure 1 is a pie chart created with Microsoft PowerPoint that illustrates the percentage of time day cares devote to various activities during a 9-hour day.

The following are options that can be considered when creating a pie chart:

1. Each pie sector has a color and/or pattern that distinguishes it from other pie sectors.
2. The smallest pie sectors must be distinguishable.

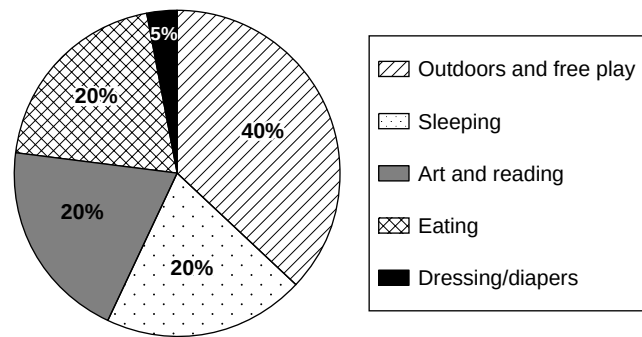


Figure 1 Percentage of Activities Pursued in Day Care

3. The largest pie sector starts at the 12 o’clock position.
4. A legend is provided to the right of the pie chart.
5. The percentages can be included either inside or outside each pie sector. They may be placed outside if the small sectors are not large enough to accommodate their size.

—Adelheid A. M. Nicol

See also Bar Chart

Further Reading

- Hollands, J. G., & Spence, I. (2001). The discrimination of graphical elements. *Applied Cognitive Psychology*, 15, 413–431.
- Rangecroft, M. (2003). As easy as pie. *Behaviour and Information Technology*, 22, 421–426.
- Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 5, 61–77.

PIERS-HARRIS CHILDREN’S SELF-CONCEPT SCALE

The Piers-Harris Children’s Self-Concept Scale (PHCSCS), published by Psychological Assessment Resources, is also titled “The Way I Feel About Myself.” It is a self-reported instrument that was designed to be used in research as well as in clinical and educational settings. Thus, the purpose of the PHCSCS is to measure conscious self-perceptions in children rather than drawing inferences from perceptions of other people

or behavioral observations. The PHCSCS is an important contribution to the field of self-concept's measuring instruments, because its items portray global as well as particular areas of self-concept prior to the current focal point of multidimensionality.

The test was conceived as a unidimensional measurement device of children's self-concept in connection with six areas of their everyday functioning: Behavior (BEH), Intellectual and School Status (INT), Physical Appearance and Attributes (PHY), Anxiety (ANX), Popularity (POP), and Happiness and Satisfaction (HAP). However, the multidimensional status of the scale has been pointed out in the most recent research. The PHCSCS includes two additional scales, the Inconsistency Index and the Response Bias Index.

The original scale contained 80 items, but a second edition consists of 60 items. Both are answered by a dichotomous "yes" or "no" response (e.g., "I am a happy person . . . yes no"). All the domains of the scale have the same names in both editions, except for two: Behavior has changed to Behavioral Adjustment (BEH), and Anxiety has changed to Freedom from Anxiety (FRE).

The theoretical rationale of its construction is based on the notion that the self-concept expresses both an overall aspect of self and special perceptions that are formed through interaction with other people. More specifically, the reasoning is derived from the following assumptions: (a) The self-concept has a phenomenological nature, (b) the self-concept has global and social elements, (c) the self-concept is relatively stable, (d) the self-concept has appraisable and descriptive elements, (e) the self-concept is affected by developmental considerations, and (f) the self-concept is organized and plays a leading part in motivation.

The PHCSCS measures the self-concept of children ages 8–18 years (Grades 4 through 12), and the test taker will need 30 minutes to complete it. The test can be administered individually or in groups. The scale yields a total score and separate scale scores, both of which are converted to stanines, percentile ranks, and *T* scores.

The PHCSCS has adequate internal consistency reliability, although its test-retest reliability is low and its construct validity is modest.

The instrument is useful for screening purposes in clinical, educational, and research settings.

—Demetrios S. Alexopoulos

Further Reading

- Bracken, B. A. (1996). *Handbook of self-concept: Developmental, social, and clinical considerations*. New York: Wiley.
- Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation*. Washington, DC: American Psychological Association.

POISSON, SIMÉON DENIS (1781–1840)

Siméon Denis Poisson was born in Pithiviers, France. In 1825, he was awarded the title Baron, and in 1837, Peer of France.

The story is told that when Poisson was a child, his nanny once suspended him from a nail to protect him from animals prior to leaving him to conduct household errands. In swinging back and forth to release himself, he became interested in the properties of the pendulum, which he later studied extensively.

His mathematical talents were first noticed by Adrien-Marie Legendre (1752–1833) when Poisson was only 17. With Legendre's support, Poisson published a paper on finite differences in *Recueil de Savants Étranges* when he was only 18.

Poisson graduated from École Polytechnique. His teachers were Joseph Louis Lagrange (1736–1813) and Pierre Simon Laplace (1749–1827). He accepted a position at Polytechnique and rose through the ranks as Lecturer (Deputy Professor) in 1802, Professor (1806), and Examiner (1816). In 1808, he gained an appointment as Astronomer at Bureau des Longitudes, and he obtained a joint appointment as Chairman of Mechanics at the Faculté des Sciences in 1809. He was elected to the physics section of the Institut National des Sciences et des Arts (Académie des Sciences) in 1812. He became an Examiner for the École Militaire in 1815.

He published about 400 works, most of which appeared in *Journal de l'École Polytechnique*, *Mémoires*

de l'Académie, and *Connaissance des Temps*. Poisson published in applied subjects, such as heat, celestial mechanics, and magnetism. His name is associated with the Poisson integral (potential theory), Poisson ratio (elasticity), and Poisson constant (electricity). Among his most important works are the *Traité de Mécanique*, with Volume 1 published in 1811 and Volume 2 published in 1833. This masterpiece is remarkable considering Poisson was never involved in conducting experiments.

He also published extensively in pure mathematics. His name is associated with Poisson brackets (differential equations). His major work in probability theory was *Recherches sur la Probabilité de Jugements*, published in 1837, in which he coined the phrase “law of large numbers,” derived from the work of Bernoulli. The major implication is that the binomial distribution tends toward the normal distribution as the number of trials increases, but as the number of successes decreases, the limit becomes the Poisson distribution.

Poisson was a gifted mathematician and teacher. His students included Michel Chasles (1793–1880), Gustav Peter Lejeune Dirichlet (1805–1859), and Joseph Liouville (1808–1882), who led the next generation's mathematicians.

—Shlomo S. Sawilowsky

Further Reading

- Ball, W. W. R. (1960). *A short account of the history of mathematics*. New York: Dover.
- Bover, C. B. (1991). *A history of mathematics* (2nd ed.). New York: Wiley.
- Cajori, F. (1919). *A history of mathematics* (2nd ed.). New York: Macmillan.
- Sawilowsky, S. S. (2004). A conversation with R. Clifford Blair on the occasion of his retirement. *Journal of Modern Applied Statistical Methods*, 3(2), 518–566.

POISSON DISTRIBUTION

A genesis of Poisson distribution with a misnomer-clature is interesting and intriguing. It was first introduced by de Movire (1718) rather than French

probabilist Siméon D. Poisson (1837), although the distribution is named Poisson. The Poisson distribution has been frequently employed to explain uncertainty in count data such as radioactive decay, traffic congestion, molecular motions, and so on, as long as it is about rarity. For an example, the number of incorrect answers in a series of questions in item response theory is considered to follow a Poisson distribution by psychologists.

For a chance mechanism to be governed and explained by a Poisson distribution, three assumptions should be validated: (a) The chance of an event occurring is proportional to the size of the time interval, which is usually infinitely small; (b) the chance of two or more events occurring together in that smaller time interval is slim; and (c) what happens in one time interval is stochastically independent of what happens in any other time interval. The probability mass function of Poisson distribution is

$$p(y) = e^{-\lambda} \lambda^y / y!,$$

where $y = 0, 1, 2, \dots$, a collection of observables in the sample space $0 < \lambda < \infty$ constitutes parametric space. The factorial moments $E[Y(Y-1) \dots (Y-k+1)]$ of order k are simply λ^k . Both Poisson events $Y = Y_{mode}$ and $Y = Y_{mode} - 1$ are equally probable when the parameter λ is an integer and equal to Y_{mode} . Otherwise, only the event Y_{mode} is the most probable event.

The Poisson distribution is a member of the linear exponential family. The Poisson probability model for counts data is popularly used to describe rare events, arrival patterns in queuing systems, particle physics, number of cancerous cells, reliability theory, risk, insurance, toxicology, bacteriology, number of accidents, number of epileptic seizures, and number of cholera cases in an epidemic, among others.

A unique property of Poisson distribution is the equality of mean and variance. That is, $\text{var}(Y) = \lambda = E(Y)$. In reality, this unique Poisson property is falsified by Poisson-like data. To match such an anomaly, several extensions of Poisson distribution have been

suggested and used. One such extension is incidence rate restricted Poisson distribution, which was introduced by Shanmugam. However, a square root transformation $\sqrt{Y}$ stabilizes the Poisson variability. When the Poisson parameter, λ , in repeated studies is noticed to be stochastically fluctuating and following a gamma probability density curve, the Poisson random variable, Y , is convoluted with a gamma density curve, and it yields an unconditional probability distribution called an inverse binomial distribution. The zero truncated inverse binomial distribution, with its scale parameter becoming negligible, approaches what is called *logarithmic series distribution*. Ecologists, including the well-known R. A. Fisher, applied logarithmic series distribution to illustrate the diversity of species on earth.

The sum $Y_1 + Y_2 + \dots + Y_n$ of two or more ($n \geq 2$) independent Poisson random variables follows a Poisson distribution with parameter equal to the sum of the parameters, $\sum_{i=1}^n \lambda_i$. Also, the conditional distribution of any one $Y_i = y_i$, given the sum $\sum_{i=1}^n Y_i = t$ of Poisson observations, follows a binomial distribution

$$p(y_i|t, n) = \binom{t}{y_i} \left(\frac{\lambda_i}{\sum_{i=1}^n \lambda_i} \right)^{y_i} \left(1 - \frac{\lambda_i}{\sum_{i=1}^n \lambda_i} \right)^{t-y_i},$$

where $y_i = 0, 1, 2, \dots, t$. Using this property, Shanmugam devised a simple and powerful test statistic to verify whether collected observations have come from an intended Poisson population or a size-biased Poisson population. Actually sampled population sometimes differs from the intended population. In a random sampling, every unit in the population should have an equal chance of being observed. Instead, the unit with a larger value has a greater chance of being observed in the sample, and such imbalanced importance of observing is recognized as size-biased sampling. The size-biased data are realizations of a not intended but actually sampled population.

Whether the Poisson parameters are equal (that is, $\lambda_1 = \lambda_2 = \dots = \lambda_n$) is tested using what is known as *dispersion test statistic* $D = \sum_{i=1}^n (y_i - \bar{y})^2 / \bar{y}$, which follows a chi-squared distribution with $n - 1$ degrees

of freedom. The maximum likelihood (also, moment) estimator of the Poisson parameter is $\bar{\lambda} = \bar{y}$. When $\lambda_1 = \lambda_2 = \dots = \lambda_n$ is true, the sum $\sum_{i=1}^n Y_i = t$ is called *complete sufficient statistic* because the conditional (binomial) distribution $p(y_i|t, n)$ is free of parameter.

Suppose that Poisson distributed counts Y_{11} , Y_{12} , Y_{21} , and Y_{22} with their parameters λ_{11} , λ_{12} , λ_{21} , and λ_{22} constitute entries in a 2×2 categorical table in case-control studies. Then, the random variable Y_{11} for given sums $Y_{11} + Y_{12} = r_1$ and $Y_{11} + Y_{21} = c_1$ follows a noncentral, hypergeometric distribution with noncentral parameter $\delta = \lambda_{11}\lambda_{12}\lambda_{21}\lambda_{22}$.

The mean and variance are equal only in Poisson distribution. Using this property, several characterizations of Poisson distribution within a specified class of distributions have been done. This property is also used as a baseline value to identify and explain the occurrence of a phenomenon called *overdispersion* in data.

Poisson is an extreme case of a binomial chance mechanism under extremely large sample size when the chance of an event is slim in each case. In a random sample of n persons from a neighborhood, the number (Y) of persons among them might have a specific disease if the chance of contracting the disease is $0 < p < 1$. That is, the binomial distribution

$$\Pr[Y = y|n, p] = \binom{n}{y} p^y (1-p)^{n-y},$$

$$y = 0, 1, 2, \dots, n,$$

converges to Poisson probability mass function as shown below.

$$\begin{aligned} \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0 \\ np = \lambda}} \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\ &= \frac{\lambda^y}{y!} \left[\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n} \right) \right]^n \left[\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n} \right) \right]^y \\ &\quad \left[\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n} \right) \dots \left(1 - \frac{y-1}{n} \right) \right] \\ &= \frac{e^{-\lambda} \lambda^y}{y!} \end{aligned}$$

Likewise, the inverse binomial distribution

$$\Pr[Y = y|k, p] = \binom{k+y-1}{y} p^k (1-p)^y,$$

$$y = 0, 1, 2, \dots, \infty,$$

converges to Poisson distribution when the so-called convolution parameter k is large. In a health screening, suppose a clinical researcher is engaged in searching for persons in a general population until a collection of k persons with a specified illness is identified.

The coefficient of variation (CV) is the ratio σ/μ of the standard deviation to the mean. The CV is scale free and illustrates the variability in the data as a percentage of its mean value. Applied scientists are fond of CV to compare several populations. For examples, CV portrays the effect of a gene in molecular studies, financial risk in stock market studies, income inequality in economics, and ecological improvement in aquatic systems. In Poisson distribution, the CV is $\lambda^{-1/2}$. Poisson and inverse binomial distributions are alternatively used to fit count data, but a comparison of their performances in terms of CV becomes difficult because their domain of possibilities is different in the scale. The CV of inverse binomial and Poisson distributions are respectively $(0, 1)$ and $(0, \infty)$. For the sake of comparing the performances of Poisson and inverse binomial distributions, Shanmugam and Singh introduced a new way of defining CV, and it is $e^{-\lambda}(1 - e^{-\lambda})$ for the Poisson distribution. The CV for the inverse binomial distribution is $(1-p)^{-k}[1 - (1-p)^{k+1}]$. Notice that the domain of Shanmugam and Singh's CV for Poisson and inverse binomial distributions are $(0, 1)$, making the comparison easier and meaningful.

For the collected Poisson distributed data on Y and predictor variables x , a Poisson regression or so-called log linear regression can be built with the relation $\lambda = \exp(\beta_0 + \beta_1 x)$, which is called *link function*. An interpretation of this link function is that when the predictor x increases by one unit, the count on Y is expected to increase by an amount e^{β_1} .

When the expected value of Poisson observation, $Y(t)$, in time interval of length t is a time-varying rate $\lambda(t)$, then it is called *nonhomogeneous Poisson*

process, and $\lambda(\tau) = \int_0^{\tau} \lambda(t)dt$ is called *Poisson intensity function*. Suppose for the i th sampling unit, there is a relation like $\lambda_i(t) = \lambda_0 \exp(\sum_{j=1}^k \beta_j x_j)$ with covariates $x_1, x_2, \dots, x_k$, then λ_0 is called the *baseline value* and the corresponding $\lambda_0[g(\tau)]$ is the *baseline intensity function*. Special cases include *exponential*, *Weibull*, or *extreme-value intensity functions* if $h(\tau) = \tau, \tau^q, e^{\alpha\tau}$, respectively.

Like in several discrete distributions of the mean exponential family, the standardized Poisson variate

$$Z = \frac{Y - E(Y)}{\sqrt{\text{var}(Y)}} = \frac{Y - \lambda}{\sqrt{\lambda}}$$

also approaches the standard Gaussian distribution. The cumulative distribution function $\Pr[Y \leq c]$ is simply the chi-squared survival function $\Pr[\chi_{2(c+1)}^2 > 2\lambda]$ with $2(c+1)$ degrees of freedom. Based on the property

$$\ln(y! \Pr[Y = y]) = -\lambda + y \ln \lambda,$$

D. C. Hoaglin and his colleagues devised a graphical technique to verify whether collected data are from a Poisson population. The data are from a Poisson population if the plot of

$$\ln\left(y! \frac{\# \text{ observations} = y}{\text{SampleSize}}\right)$$

versus y is a linear line with intercept and slope equal to $-\lambda$ and $\ln \lambda$. The so-called damage model establishes that unless the damaged part D and the undamaged part $Y - D$ follow independently a Poisson distribution with parameters λp and $\lambda(1-p)$ respectively, the complete observation $Y = D + \text{undamaged}$ follows a Poisson distribution with parameter λ where p denotes the chance for an observation to be damaged. For an example, note that Y and D might be the number of cancerous and successfully eliminated cancerous cells in a laser therapy.

In spite of being the natural choice for data on rare events, the Poisson distribution has failed many times. Shanmugam stated that there is only a 4% chance that the chance mechanism regarding the

number of accidents per day in a location near Virginia State will be a Poisson. With respect to the number of chromatid aberrations in gene mutation analysis, the performance of Poisson distribution has been reported to be worse, even though the data are about rareness. Where does the fallacy occur? Is it due to the lack of equality of mean and variance? To combat such mistrust, what is commonly done in the Bayesian approach is that the Poisson parameter is assumed to follow a probability distribution. This approach results in an inverse binomial distribution, but we part company with Poissonness. Instead, Shanmugam recommends using what is known as *incidence rate restricted Poisson distribution* with a probability mass function

$$\Pr[Y = y|\lambda, \beta] = \left(1 + \frac{y}{\beta}\right)^{y-1} (\lambda e^{-\lambda/\beta})^y / y! e^{\lambda},$$

where the incidence parameter λ is restricted by β on the upper side and $y = 0, 1, 2, \dots$. The moment estimate of mean and variance of the incidence rate restricted Poisson are respectively $\hat{\lambda} = \bar{y}^{3/2}/s$ and $\hat{\beta} = \hat{\lambda}/(1-(\bar{y}/s^2)^{1/2})$ where $\bar{y}$ and s^2 are sample mean and variance, respectively. When $\hat{\beta} = \infty$, or, equivalently, $s^2 = \bar{y}$, the above probability mass function of the incidence rate restricted Poisson distribution simplifies to Poisson distribution. This implies that the usual Poisson distribution is appropriate when the incidence parameter is unrestricted. The size-biased incidence rate restricted Poisson distribution with probability mass function

$$\begin{aligned} \Pr[Y = y|\lambda, \beta] &= \frac{y \Pr[Y = y|\lambda, \beta]}{E(Y)} \\ &= (1 - \lambda/\beta) e^{-\lambda(1+\lambda/\beta)} [\lambda e^{-\lambda/\beta} (1 + y/\beta)]^{y-1} / (y-1)! \end{aligned}$$

was found to fit well with international terrorism data by Shanmugam, where $y = 0, 1, 2, \dots$.

When both the probability mass and the mean functions can be expressed in an exponential format, it has been defined by Shanmugam to be *mean exponential family* (MEF) distributions. A Poisson distribution is a member of the MEF. Using a characterization of MEF, Shanmugam devised a

goodness-of-fit test for an unbiased and a size-biased Poisson sample. That is, a nonnegative random sample $y_1, y_2, \dots, y_n$ of size n could be considered to have come from an intended (or size-biased) Poisson population with $100(1 - \alpha)\%$ confidence level if the standardized score

$$z = \left| \left[\sum_{i=1}^n y_i^2 - n\bar{y}(\bar{y} + 1) \right] / \bar{y}\sqrt{2(n-1)} \right| \text{ or,}$$

$$z = \left| \left[\sum_{i=1}^{n^*} y_i^2 - n^*(\bar{y}^2 + 1) \right] / (1 + \bar{y}^*)\sqrt{\bar{y}^*(n^* - 1)} \right|$$

is less than or equal to $z_{\alpha/2}$, respectively, where $n^* = n - \# \text{ zero observations}$ and $\bar{y}^* = \# \text{ nonzero observations over } n^*$.

—Ramalingam Shanmugam

Further Reading

- Haight, F. A. (1967). *Handbook of the Poisson distribution*. New York: Wiley.
- Kagan, A. M., Linnik, Y. V., & Rao, C. R. (1973). *Characterization problems in mathematical statistics*. New York: Wiley.
- Moivre, A. de (1718). *The doctrine of chance: Or, a method of calculating the probability of events in play*. London: Pearson.
- Poisson, S. D. (1837). *Recherches sur la Probabilite des Jugements en Matiere Criminelle et en Matiere Civile, Precedees des Regles Generales du Calcul des Probabilities*. Paris: Bachelier, Imprimeur-Libraire pour les Mathematiques, la Physique, etc.
- Shanmugam, R. (1989). Asymptotic homogeneity tests for mean exponential family distribution. *Journal of Statistical Planning and Inference*, 23, 227–241.
- Shanmugam, R. (1991). Incidence rate restricted Poissonness. *Sankhya*, B, 63, 191–201.
- Shanmugam, R. (1992). Fractions in Morris' quadratic variance natural exponential family of distributions and a lack of consensus in sample size determination. *Statistics and Probability Letters*, 15, 20–25.
- Shanmugam, R. (1993). Size-biased incidence rate restricted Poissonness and its application in international terrorism. *Applications of Management Science: Public Policy Application of Management Sciences*, 7, 41–49.
- Shanmugam, R. (1995). Goodness of fit test for length-biased data with discussion on prevalence, sensitivity and specificity

of length bias. *Journal of Statistical Planning and Inference*, 48, 277–280.

Shanmugam, R. (2003). Index of variation with illustration using molecular data. *Communications in Statistics*, 32, 509–517.

PORTFOLIO ASSESSMENT

A portfolio is a systematic, purposeful, and ongoing collection of students' work that exhibits their efforts and achievements in one or more skill domains. The content of a portfolio is created in response to specific instructional objectives and evaluated in relation to those objectives. Although teacher guidance and support are provided, portfolio development is primarily the responsibility of the learner. Portfolio assessment is not a single assessment method, but rather a process of collecting, compiling, and maintaining information from multiple sources to evaluate students' learning and development. This type of assessment places an emphasis on what the learner can do, not on identifying deficits or comparing performance and work products with those of other students. Portfolio assessment is also an effective communication system for conveying information to students, other teachers, and parents about students' progress and unique accomplishments.

Although approaches to portfolio development and assessment may vary, all share several key characteristics. Portfolios reflect explicit learner outcomes, identified in the curriculum and/or by teachers, that students are expected to achieve; they focus on students' performance-based learning experiences as well as their acquisition of knowledge and skills; they contain samples of work that range over an entire grading period, rather than a single point in time (e.g., a single test score); and they include a variety of different types and formats of evidence (e.g., written, oral, graphic, media-based).

Benefits of Portfolio Assessment

Portfolio assessment is an alternative to the use of traditional grading systems and standardized

testing procedures, which may not provide teachers with a clear understanding of students' accomplishments and, often, are not directly linked to instruction. The development of portfolio assessment resulted from concerns surrounding traditional assessment procedures and addressed the need for an assessment method capable of demonstrating what individual students know and what they can accomplish. Traditional assessment practices measure students' ability at one point in time, are usually conducted outside the context of classroom instruction, and may not capture the full range of students' abilities. In contrast, portfolio assessment measures students' ability over time, is embedded within instructional activities, and captures multiple facets of students' performance across diverse skill areas.

There are multiple advantages to portfolio assessment, especially for instructional planning and student progress monitoring. First, portfolio assessment links assessment information directly to teaching and instructional outcomes. Portfolio artifacts are authentic examples of classroom work, not divorced from instructional activities as some test items can be. Portfolio assessment is viewed as *embedded assessment*, that is, assessment in which the assessment tasks are actually part of instruction. As such, portfolio assessment is an effective method to bring assessment in line with instructional goals and provides teachers with valuable information on which to base their instructional decisions.

Second, because the products and evidence are collected over time, portfolios serve as a record of students' growth and progress. Portfolio assessment is longitudinal and continuous, providing both formative (ongoing) and summative (culminating) opportunities for monitoring students' progress toward achieving essential learning outcomes. It provides a comprehensive body of student work that is used to evaluate performance over time.

Third, using portfolio assessment enables teachers to broaden their curriculum to include areas they traditionally could not assess with standardized testing or classroom exams. Portfolio assessment complements teachers' use of innovative instructional

strategies centered around the use of teamwork, group projects, and applied learning. Portfolio assessment is also compatible with individualized instruction and teaching strategies that focus on unique learning styles. Rather than teaching “to the test” and gearing instruction toward achieving high test scores, portfolio assessment encourages teaching and learning to help students acquire and apply new knowledge.

Fourth, a portfolio is a form of assessment that students complete together with teachers. An important dimension of portfolio assessment is that it encourages teacher-student collaboration and actively involves students in the process of assessment. Students are involved in choosing and justifying the portfolio pieces to be included. Portfolio assessment enables teachers to get to know students and promotes joint goal setting and negotiation of grades. It affords opportunities for students and teachers to discuss learning goals and progress toward these goals through both structured and informal conferences.

Fifth, a portfolio enables students to demonstrate their highest quality work, without the pressure of high-stakes testing or time constraints. Portfolios are multidimensional and include many different types of artifacts that demonstrate various aspects of students’ learning and a range of skills. Students can include work at their own level and related to their personal interests; portfolios accommodate different learning styles and allow for demonstration of individual learner strengths. Furthermore, students are empowered to select and justify their portfolio entries, monitor their own progress, and establish learning goals. The empowerment of students to demonstrate their achievements and personal accomplishments through portfolio assessment has been shown to enhance student motivation. Few traditional assessment practices afford students the opportunity to assume an integral role in their own learning and evaluation. Students feel a pride of ownership of their portfolio work, and they see the personal and academic relevance of the work they have completed. Most importantly, students value themselves as learners as they proceed through the

portfolio process of selecting work and reflecting on each piece.

Implementation of Portfolio Assessment

Implementation of portfolio assessment involves two basic components, portfolio development and portfolio evaluation. The development of a portfolio occurs across three distinct phases. During Phase 1, Organization and Planning, students and teachers collaborate to organize and plan the development of a student’s portfolio. Questions for students to address during this planning phase include the following: How do I select work samples to reflect what I am learning? How do I organize and present the work I have collected? How will I maintain and store my portfolio? Phase 2, Collection, involves the collection and compilation of work products that reflect the student’s accomplishments and attainment of educational goals. Portfolios may consist of a variety of written, oral, graphic, or media content that demonstrate progress over time, including notes, videotapes of presentations, photographs, logs, written summaries, and samples of work. Students may also be evaluated on projects or assignments completed in pairs or small groups of students. Phase 3, Reflection, is a critical feature of portfolio development. During this phase, students engage in self-reflection on the learning process and their developing knowledge and skills. Students are required to evaluate their own progress and the quality of their work in relation to targeted goals or competencies. By reflecting on their own learning, students begin to identify their strengths and weaknesses. Identified weaknesses become goals for improvement and future learning. Student reflections may take the form of learning logs, reflective journals, or self-ratings. When possible, teacher and parent reflections on the student’s products and learning are also included in a portfolio.

For effective evaluation of a portfolio to occur, teachers must have scoring criteria in place that are clearly communicated and understood by students. The criteria for assessment of the portfolio contents is articulated at the outset of the process. Portfolio

evaluation is based on multiple criteria, such as the student's growth and development in relationship to curriculum expectations and goals; his or her understanding of key concepts; the comprehensiveness, accuracy, and relevance of work products; diversity of entries; and thoughtfulness of self-reflections and justification of the selected artifacts. Scoring keys, rules, or rubrics are designed for portfolios and shared with students prior to evaluation. Letter grades, combined with evaluative feedback, may be assigned to portfolios based on the scoring rubric. Finally, a form of oral interview or discussion between the teacher and student is often included as part of the portfolio evaluation process.

Conclusion

Portfolio assessment fits well with the growing trend in education toward monitoring and promoting all students' progress toward achievement of targeted outcomes. Portfolio assessment improves upon traditional student testing practices by revealing a range of skills and understandings. Portfolio assessment supports instructional goals, reflects students' progress and growth over a period of time, and encourages students' self-assessment and self-directed learning.

—Maribeth Gettinger

Further Reading

- Barton, J., & Collins, A. (Eds.). (1997). *Portfolio assessment: A handbook for educators*. Menlo Park, CA: Addison-Wesley.
- Cole, D. J., Ryan, C. W., & Kick, F. (1995). *Portfolios across the curriculum and beyond*. Thousand Oaks, CA: Corwin.
- Grace, C., Shores, E., & Charner, K. (Eds.). (1998). *The portfolio book: A step-by-step guide for teachers*. Beltsville, MD: Gryphon House.
- MacDonald, S. (1997). *The portfolio and its use: A road map for assessment*. Beltsville, MD: Gryphon House.
- Seidel, S., Walters, J., Kirby, E., Olff, N., & Powell, K. (1998). *Portfolio practices: Thinking through the assessment of children's work*. Washington, DC: National Education Association.
- Shaklee, B. D., Barbour, N. E., Ambrose, R., & Hansford, S. J. (1997). *Designing and using portfolios*. Boston: Allyn & Bacon.

Wingograd, P., & Jones, D. L. (1992). The use of portfolios in performance assessment. *New Directions for Educational Reform*, 1(2), 37–50.

Portfolio assessment and implementation checklists resources: <http://www.pampetty.com/assessment.htm>

Pportfolio assessment teacher resources: http://www.phschool.com/professional_development/assessment/portfolio_based_assess.html

POST HOC COMPARISONS

Post hoc comparisons among sample means from three or more groups are generally performed only after obtaining a significant omnibus *F* when we use an ANOVA. After we find the various means are not all equal, the second step is using post hoc comparisons to find out exactly which means are significantly different from which other ones. In contrast to a priori comparisons, which are chosen before the data are collected, post hoc comparisons are tested after the researcher has collected the data.

Post hoc comparisons include pairwise comparisons and nonpairwise comparisons. Pairwise comparisons compare two sample means at a time, whereas nonpairwise comparisons compare more than two sample means at a time.

The main post hoc comparison procedures include Scheffé procedure and Tukey HSD (honestly significant difference) procedure. The Scheffé procedure allows for a comparison of all possible paired comparisons and complex comparisons between combined means.

The formula for the Scheffé test is as follows:

$$F_c = \frac{(\sum C_j \bar{X}_j)^2}{(MS_W) \left(\sum \frac{C_j^2}{n_j} \right)},$$

where

Σ , the Greek letter sigma, is the summation sign;

C_j is the coefficient used for any group;

$\bar{X}_j$ is the mean of the corresponding group;

MS_w is the mean square within from the analysis of variance.

The Tukey HSD procedure only allows for a comparison of the possible pairs of means. The formula for the Tukey HSD test is as follows:

$$HSD = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(MS_w/2)(1/n_1 + 1/n_2)}},$$

where

$\bar{X}_1$ and $\bar{X}_2$ are two sample means that are needed for comparison,

MS_w is the mean square within from the analysis of variance,

n_1 is the number of scores of Group 1,

n_2 is the number of scores of Group 2.

For example, the data set in Table 1 consists of 10 cases with two variables, Group and Test Score.

To produce the post hoc comparison, follow these steps:

1. Compute the sample mean for each group.

	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>
Mean	9.70	10.70	9.80	7.10

2. Check the assumption of equal variance.

<i>Levene Statistic</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>
1.429	3	36	.250

To test the null hypothesis that groups come from populations with the same variance, the Levene test is produced. The observed significance level is larger than .05. The null hypothesis is not rejected. Four groups come from populations with the same variance.

3. Produce ANOVA to get the omnibus F test.

The null hypothesis of the F test is that all the population means are the same.

Table 1 Post Hoc Comparison Data Set

<i>Group</i>	<i>Test Score</i>
1	10
1	9
1	13
1	13
1	9
1	8
1	5
1	10
1	9
1	11
2	13
2	12
2	10
2	11
2	13
2	14
2	12
2	9
2	8
2	5
3	11
3	10
3	10
3	12
3	10
3	8
3	7
3	11
3	10
3	9
4	7
4	9
4	5
4	7
4	6
4	5
4	6
4	8
4	9
4	9

The alternative hypothesis is that the population means are not all equal.

Use a .05 significant level.

	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	72.075	3	24.025	5.382	.004
Within Groups	160.700	36	4.464		
Total	232.775	39			

According to the results of the F test, $F(3,36) = 5.382$, $p = .004$, $p < .05$, so we reject the null hypothesis, and the four population means are not all equal.

4. Use post hoc comparison to determine which means are significantly different from each other.

The results of the Scheffé procedure show that only Group 2 and Group 4 are significantly different from

each other, whereas the results of the Tukey HSD procedure show that Group 4 is significantly different from Group 1, Group 2, and Group 3. The reason is that the Scheffé test is more considerate than the Tukey HSD test in holding Type I error low.

The SPSS output is shown in Tables 2 and 3.

—Bixiang Ren

Table 2 Test of Homogeneity of Variances

testscore			
<i>Levene</i>			
<i>Statistic</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>
1.429	3	36	.250

Table 3 Post Hoc Tests: Multiple Comparisons

Dependent Variable: testscore

	<i>(I)group</i>	<i>(J)group</i>	<i>Mean Difference (I - J)</i>	<i>Std. Error</i>	<i>Sig.</i>
Tukey HSD	1.00	2.00	−1.00000	.94487	.717
		3.00	−.10000	.94487	1.000
		4.00	2.60000(*)	.94487	.044
	2.00	1.00	1.00000	.94487	.717
		3.00	.90000	.94487	.777
		4.00	3.60000(*)	.94487	.003
	3.00	1.00	.10000	.94487	1.000
		2.00	−.90000	.94487	.777
		4.00	2.70000(*)	.94487	.034
	4.00	1.00	−2.60000(*)	.94487	.044
		2.00	−3.60000(*)	.94487	.003
		3.00	−2.70000(*)	.94487	.034
Scheffé	1.00	2.00	−1.00000	.94487	.773
		3.00	−.10000	.94487	1.000
		4.00	2.60000	.94487	.073
	2.00	1.00	1.00000	.94487	.773
		3.00	.90000	.94487	.823
		4.00	3.60000(*)	.94487	.006
	3.00	1.00	.10000	.94487	1.000
		2.00	−.90000	.94487	.823
		4.00	2.70000	.94487	.059
	4.00	1.00	−2.60000	.94487	.073
		2.00	−3.60000(*)	.94487	.006
		3.00	−2.70000	.94487	.059

* The mean difference is significant at the .05 level.

Further Reading

Poisson distribution definitions and illustrations: http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html#poisdistn

Poisson distribution generating applet: <http://www.math.csusb.edu/faculty/stanton/probstat/poisson.html>

POSTERIOR DISTRIBUTION

The term *posterior distribution* refers to a probability density function for an unobserved or latent variable, θ , based on the observed data, x , used in Bayesian statistical analysis. In Bayesian data analysis, inferences are made on the basis of the posterior distributions of parameters. Bayes' theorem is generally phrased in terms of distributions of observed and unobserved variables:

$$f(\theta | x) = \frac{f(x | \theta)f(\theta)}{f(x)}.$$

By using this formulation of Bayes' theorem, statisticians are able to make inferences about some parameter of interest, θ , given the observed data, x . Therefore, under the framework of a Bayesian data analysis, statistical inferences are based on a quantity that is of direct interest to the analyst (i.e., θ), not some proxy for that quantity of interest (i.e., the data, x).

In order to estimate the posterior distribution and make inferences about it, three pieces of information are required: (a) estimation of the function, $f(x | \theta)$, often termed the "likelihood function," which represents a statistical model that has been fit to the distribution of observed data, x , given the underlying parameter, θ ; (b) estimation of $f(\theta)$, referred to as the "prior distribution," which represents either the empirical or expected distribution of the parameter, θ , in the population, and (c) estimation of $f(x)$, which represents the empirical distribution of the observed data, x .

Inferences from the posterior distribution are typically made by determining point estimates for θ , either by finding the mean of the posterior distribution (referred to as an "expected a posteriori" estimate) or by determining the mode of the posterior distribution

(referred to as a "modal a posteriori" estimate). The standard error of θ is determined by estimating the standard deviation of the posterior distribution.

—William P. Skorupski

See also Bayesian Statistics; Prior Distribution

Further Reading

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

PREDICTIVE VALIDITY

There are several different types of validity, all of which are used to establish a test's "truthfulness," or whether the test does what it is supposed to do (the main claim of validity). Among the many different kinds of criterion validity, a subset is concurrent and predictive validity.

Predictive validity allows the measurement specialist to judge how well a test predicts future performance. The interest is in whether the instrument can predict what it theoretically should be able to predict.

For example, the following outcomes would characterize test and measurement tools that have predictive validity.

- A math test predicts future performance in an occupation where sophisticated mathematical skills are essential, such as engineering or computer science.
- A test of spatial skills predicts future performance in an occupation where the ability to visualize in more than two dimensions is essential, such as mechanically oriented tasks and professions.
- A test of personality predicts the configuration of personal skills that might be needed to succeed in a particular profession, such as hotel management.

What these three hypothetical statements have in common is that each of the tests that is developed is predictive of a later outcome (which is the criterion) and hence, if they do their job, they have predictive validity.

Predictive validity is most often established through the simple Pearson product-moment correlation

computed between the test itself and the criterion. To then establish the predictive validity of the mathematics test mentioned above, test scores would be collected and then correlated with some criterion used to evaluate engineering skills a certain number of years later. That simple zero-order correlation indicates the strength of the association between the test and the criterion and, hence, the predictive validity of the test.

—Neil J. Salkind

See also Concurrent Validity; Criterion Validity; Validity Theory

Further Reading

Predictive validity of the Graduate Management Admissions Test report: <http://www.gmac.com/gmac/NewsCenter/Resources/FactSheetGMATValidity.htm>

Predictive validity of the Graduate Record Examinations advanced psychology test for grade performance in graduate psychology courses: http://www.findarticles.com/p/articles/mi_m0FCR/is_1_36/ai_85007765

PRESCHOOL LANGUAGE ASSESSMENT INSTRUMENT

The Preschool Language Assessment Instrument (PLAI-2), published by PRO-ED (www.proedinc.com), is a nationally standardized tool for assessing children's discourse abilities. It consists of 70 items that are classified in terms of four levels of abstraction (matching, selective analysis, reordering, and reasoning) and two modes of responding (receptive and expressive). Items from each category are interspersed to simulate the demands of classroom discourse. The test, which takes about 30 minutes to administer, is intended for children between 3 years, 0 months to 5 years, 11 months. Scoring procedures classify responses according to the levels of abstraction and modes of responses yielding scores for six subtests for 4- to 5-year-olds. Reordering and reasoning are combined for the 3-year-olds, resulting in five subtests. The inclusion of normative data permits evaluation of whether children's classroom discourse skills are developing normally. The test also permits identification

of strengths and weaknesses and documentation of change over time.

The four levels of abstraction represent increasing levels of difficulty encountered in classroom discourse. Receptive and expressive items are included in each level. Matching, the lowest level of abstraction, involves naming items (e.g., "What is this?"), pointing to named items (e.g., "Find me a cup"), following directions, or imitating. Some items require visual matching and visual memory. Selective Analysis, the next level of abstraction, requires selective attention to visual detail. The receptive items require understanding descriptions (e.g., given a picture, "show me what we use for cleaning dishes" or "find something we could eat with"), following multistep directions, and identifying differences. The expressive items involve answering a variety of *Wh* questions (e.g., *who*, *what*, *where*, *how*, *which*, and *what is happening*) in response to pictures or a brief narrative. Auditory memory, integration, and classification are required for successful responses at this level. The third level of abstraction, Reordering, requires overriding perceptual clues while selecting or identifying multiple features in a picture (e.g., "If I wanted to paint a picture, show me all the things I don't need"). This level also requires identifying similarities, defining words (e.g., "Tell me what a fork is") or solving class inclusion problems. The highest level of abstraction, Reasoning, requires predicting events under specified conditions (e.g., "Given a stack of blocks, select the picture that shows what will happen if the bottom block is removed"). The responses at this highest level specify predictions, logical justifications, and causal relations.

Raw scores for each subtest are converted into scaled standard scores, percentile ranks, descriptive ratings, and age-equivalent scores. Scaled scores for the receptive and expressive subtests are converted into an overall Discourse Ability Score. The expressive responses are also scored in terms of four levels of adequacy and six types of interfering behaviors. The authors describe these latter types of scores as measures of pragmatic aspects of communication.

An original 165-item version and a more practical 60-item experimental edition of the test were

developed in 1978. The original versions and the current revised test are based on a model of classroom discourse in which a child is required to respond to varying levels of abstraction. At the lowest level, language is connected to concrete perceptions, whereas at the highest level, language is used to reflect upon perceptions. The PLAI-2 has expanded the number of items from 60 to 70. It has added normative data representative of the 2000 U.S. population, full-color illustrations, new reliability and validity data, as well as two additional scales related to adequacy of responses and interfering behaviors. Reported data indicate that the overall Discourse Ability Score meets current minimum standards for test-retest reliability.

At the present time, published literature on the test refers to the original versions. A positive feature of the test is the focus on assessing classroom discourse skills. Issues regarding validity are discussed in Skarakis-Doyle, Miller, and Reichheld.

The manual for the PLAI-2 addresses some of the issues raised in the review.

—Jennifer R. Hsu

Further Reading

- Blank, M., Rose, S. A., & Berlin, L. J. (1978). *The language of learning: The preschool years*. New York: Grune & Stratton.
- Haynes, W. O. (1985). Review of Preschool Language Assessment Instrument, Experimental Edition. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 1190–1192). Lincoln: University of Nebraska Press.
- Skarakis-Doyle, E., Miller, L. T., & Reichheld, M. (2000). Construct validity as a foundation of evidence-based practice: The case of the Preschool Language Assessment Instrument. *Journal of Speech-Language Pathology and Audiology*, 24(4), 180–191.
- Skarakis-Doyle, E., Miller, L. T., & Reichheld, M. (2001). Construct validity as a foundation of evidence-based practice: The case of the Preschool Language Assessment Instrument: Erratum. *Journal of Speech-Language Pathology and Audiology*, 25(1), 40.
- Marion Blank biography: <http://www.laureatelearning.com/professionals602/LLSinfo/authors.html#anchor30704660>

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a multivariate analysis technique whose goal is to reduce the dimensionality of a large number of interrelated variables. It belongs to the class of projection methods and achieves its objective by calculating one or more linear combinations of the original set of maximum variance. This criterion implies that the directions of maximum variability in the data are those of interest. PCA is a fairly old technique, because its basic formulation can be traced back to the work of Pearson, whereas its usual algebraic derivation is found in the work of Hotelling. Moreover, for time series data, it corresponds to the Karhunen-Loeve decomposition of Watanabe. In its basic formulation, PCA is most suitable for numerical data; however, extensions to categorical data have been proposed in the literature, together with some generalizations suitable for data exhibiting nonlinearities. PCA has been proven to be a very successful dimension reduction technique, widely used in many scientific fields.

PCA Formulation

Data on K variables have been collected for N objects. The data are organized in an $N \times K$ matrix $\mathbf{X}$. We will denote the variables by $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K$ that correspond to the columns of the data matrix $\mathbf{x}$. It is customary to work with *standardized* versions of the variables—that is, $\text{Mean}(\mathbf{X}_k) = 0$ and $\text{Var}(\mathbf{X}_k) = 1$, $k = 1, \dots, K$, the reason being to avoid the scale (range) of the variables to affect the results.

PCA constructs linear combinations of the original variables $\mathbf{y}_j = \mathbf{X}\mathbf{w}_j$, $j = 1, \dots, K$ that have maximum variance. Collecting the $\mathbf{y}_j$ variables in a matrix $\mathbf{Y} = \mathbf{X}\mathbf{W}$, the PCA problem can be written as

$$\max_{\mathbf{W}} \text{Cov}(\mathbf{Y}) = \max_{\mathbf{W}} (\mathbf{W}^T \mathbf{R} \mathbf{W}), \quad (1)$$

where $\mathbf{R}$ denotes the correlation matrix of the data matrix $\mathbf{X}$. It is further required that $\mathbf{W}^T \mathbf{W} = \mathbf{I}$, in order to identify a unique solution. An application of the Rayleigh-Ritz theorem shows that by performing an eigenvalue decomposition of $\mathbf{R} = \mathbf{U}\mathbf{A}\mathbf{U}^T$

and setting $\mathbf{W} = \mathbf{U}$, the optimal weight matrix $\mathbf{W}$ is obtained. Hence, the j th column of $\mathbf{W}$ corresponds to the eigenvector of $\mathbf{R}$ associated with the j th largest eigenvalue.

The $\mathbf{y}_j$ variables are called the principal components. Some of their properties useful for understanding and interpreting PCA are outlined next.

- The principal components have mean zero and are uncorrelated.
- The variance of the j th principal component is given by the corresponding eigenvalue; that is, $\text{Var}(\mathbf{y}_j) = \lambda_j$.
- The correlation of the j th principal component with original variables is given by $\text{Corr}(\mathbf{y}_j, \mathbf{x}_i) = \sqrt{\lambda_j} w_{i,j}$. The latter quantity is called the *factor loading*.
- The proportion of total variance explained by each principal component is given by $\lambda_j / \sum_{j=1}^p \lambda_j$.

Remark: The above presentation took a data centric point of view. An alternative derivation could have been based on a population point, with $\mathbf{R}$ corresponding to a known population correlation matrix.

Remark: Over the years, a considerable amount of work has been devoted to deriving the probability distribution of sample principal components and their variances. The results are mostly asymptotic, and they are often based on the assumption that the original set of variables follows a multivariate normal distribution.

Remark: The derivation of PCA was based on the sample correlation matrix. However, if one decided not to standardize the data, then the optimal weights $\mathbf{W}$ correspond to the eigenvectors of the sample covariance matrix. Unfortunately, there is no straightforward mathematical relationship that links the two analyses, namely, that based on the correlation matrix and that based on the covariance matrix.

PCA is illustrated next using a data set that gives crime rates per 100,000 people for the 72 largest U.S. cities in 1994. The variables correspond to the following seven crimes: murder, rape, robbery, assault, burglary, larceny, and motor vehicle thefts (MVTs). Because of the disparity in the scale of the variables, PCA was based on the correlation matrix of the data. The first two eigenvalues accounted for 54% and 17% of the total variance, respectively.

The optimal weights $\mathbf{W}$ of the crime variables for the first two principal components are given in Table 1.

A projection of the 72 cities on the space spanned by the first two principal components is shown in Figure 1.

Looking at the weights of the first two principal components, it can be seen that the first one can be interpreted as an overall measure of crime activity. With this interpretation in mind, the cities on the right side of the plot, such as St. Louis, Atlanta, Tampa Bay, Newark, Detroit, and Miami, can be characterized as high-crime cities, whereas cities appearing on the left side, such as Virginia Beach, San Jose, Colorado, and Honolulu, can be characterized as low-crime cities. On the other hand, the second principal component distinguishes between cities with high rape and larceny incidents (and to some degree, burglaries) and cities with high murder, robbery, and MVT incidents. So, on the bottom of the picture, we find cities such as Newark; Jersey City; Philadelphia; Santa Ana; Detroit; Washington, DC; New York City; Chicago; and Long Beach characterized by relatively more murder, robbery, and MVT crimes, whereas in Oklahoma City, Corpus Christi, Tucson, and Minneapolis, rapes and larcenies are more frequent. However, one should be careful regarding how far to proceed with this interpretation. For example, it is appropriate to make such statements for Newark and Detroit, which score high on the first component as well. On the other hand, the situation is not that clear for cities like Santa Ana, whose score is fairly low on the first component. Nevertheless, the second component allows one to distinguish between Corpus Christi and Santa Ana, which score similarly on the first component. Therefore, the

Table 1 Optimal Weights (Loadings) for the Crimes for the PCA of the 1994 U.S. Crime Rates

Variable	PC1	PC2
Murder	0.370	-.339
Rape	0.249	0.466
Robbery	0.426	-.387
Assault	0.434	0.042
Burglary	0.449	0.238
Larceny	0.276	0.605
MVT	0.390	-.302

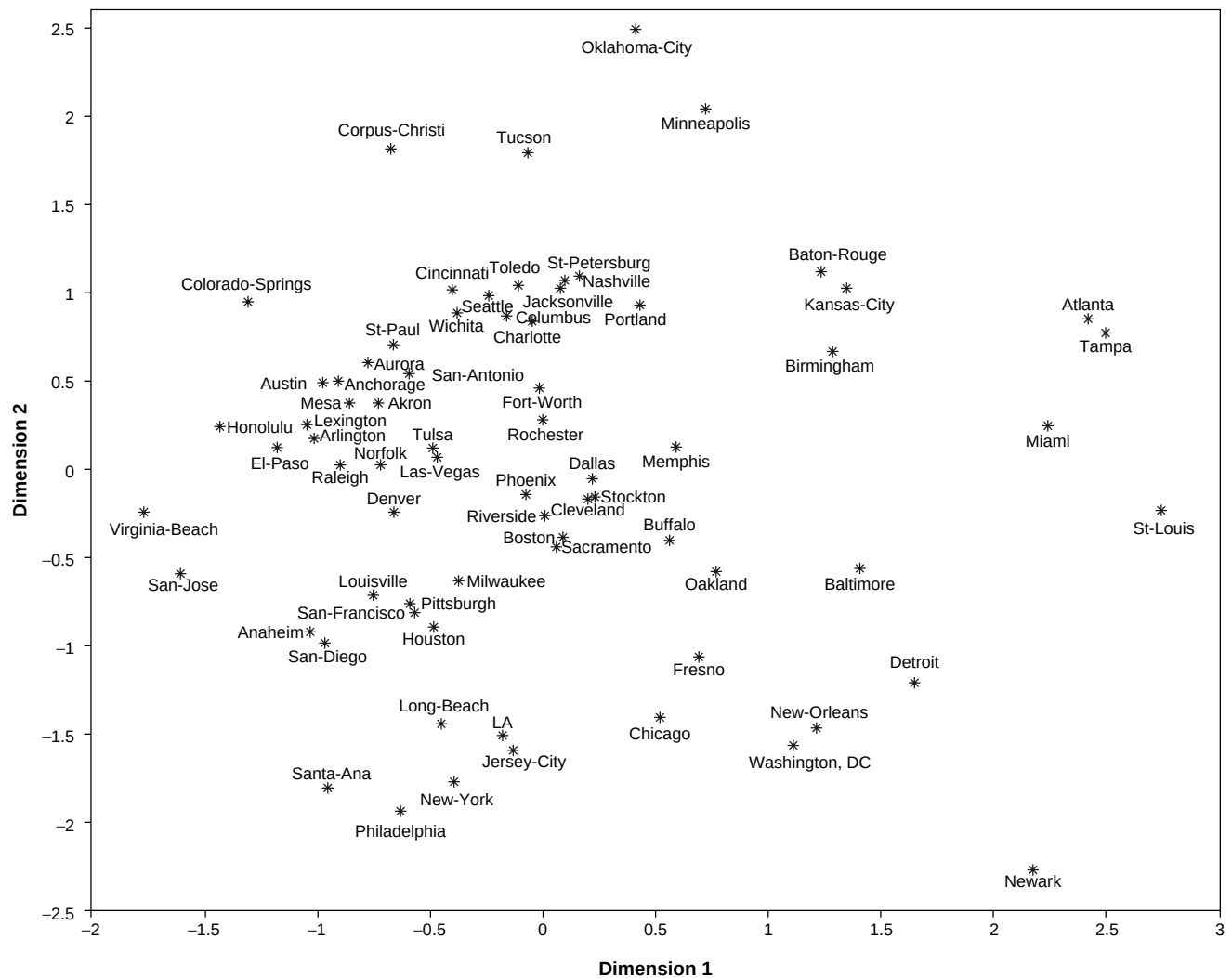


Figure 1 PCA of the 1994 U.S. Crime Rates

Source: Michailidis & de Leeuw (1998).

Note: Projection of the 72 American cities on the first two principal components.

plot suggests that there are many more rapes in Corpus Christi compared to Santa Ana or New York City, whereas the latter two cities have more murders.

Generalizations of PCA

A variation of PCA is suitable for performing dimension reduction to categorical data. The corresponding version of the technique is known as *multiple correspondence analysis*.

Other generalizations try to deal with the presence of nonlinearities in the data. Some of these approaches are principal curves, local PCA, and kernel PCA.

Hastie and Stuetzle proposed the concept of a principal curve, where the data are mapped to the closest point on the curve, or, alternatively, every point on the curve is the average of all the data points that are projected onto it (a different implementation of a centroid-like principle). The regular PCA solution produces the only straight lines possessing the above property.

Local PCA attempts to preserve the conceptual simplicity and algorithmic simplicity of PCA in the presence of nonlinearities. One implementation first uses cluster analysis to group the data and subsequently applies different principal component analyses to the various clusters.

Finally, the main idea behind kernel PCA is that nonlinearities in the original data essentially disappear if the data are projected onto some higher dimensional space, at which point regular PCA becomes an effective solution. Although this idea is contrary to the spirit of PCA as a dimension reduction technique, it has nevertheless enjoyed some success in application areas such as handwritten digit recognition. In order to make this idea computationally tractable, kernels (to some extent, they can be thought of as generalizations of a covariance function) are used, which essentially calculate inner products of the original variables. It should be noted that kernels are at the heart of support vector machines, a very popular and successful classification technique.

—George Michailidis

See also Eigendecomposition; Correspondence Analysis; Factor Analysis; Multiple Correspondence Analysis; Multiple Factor Analysis; Singular and Generalized Singular Value Decomposition; STATIS

Further Reading

- Bregler, C., & Omohundro, M. (1994). Surface learning with applications to lipreading. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems*. San Mateo, CA: Morgan Kaufman.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, UK: Wiley.
- Golub, G. H., & van Loan, C. F. (1989). *Matrix computations*. Baltimore: Johns Hopkins University Press.
- Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.
- Liu, Z. Y., & Xu, L. (2003). Topological local principal component analysis. *Neurocomputing*, 55, 739–745.
- Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13, 307–336.
- Scholkopf, B., Smola, A., & Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1299–1319.

PRIOR DISTRIBUTION

The term *prior distribution* refers to an empirical or expected probability density function used in

Bayesian statistical analysis, which represents an analyst's belief about the distribution of an unobserved parameter, θ , in the population. In Bayesian data analysis, inferences are made on the basis of estimated probability distributions for unobserved or “latent” variables based on observed data. Bayes' theorem is generally phrased in terms of distributions of observed and unobserved variables:

$$f(\theta | x) = \frac{f(x | \theta)f(\theta)}{f(x)}.$$

By using this formulation of Bayes' theorem, statisticians are able to make inferences about some parameter of interest, θ , given the observed data, x . This density function, $f(\theta | x)$, is referred to as the *posterior distribution* of θ , and it represents a probability density function for the latent variable, θ , based on the observed data, x . In this formulation, $f(x | \theta)$ is the “likelihood function,” which represents a statistical model that has been fit to the distribution of observed data, x , given the underlying parameter, θ . $f(x)$ is the empirical distribution of the observed data, x , and $f(\theta)$ is the prior distribution for the unobserved parameter, θ .

Whereas $f(x | \theta)$ and $f(x)$ are quantities estimated using observed data, the prior distribution, $f(\theta)$, is typically unobserved and therefore must be selected by the analyst based on a belief about how the parameter θ is distributed in the population. Selection of an appropriate value for a prior distribution is therefore critical to the accuracy of results from a Bayesian statistical analysis.

Bayesian procedures incorporate information from the prior distribution in order to make inferences. Generally speaking, the prior distribution for θ is an unobserved probability density function that must be estimated somehow by the analyst. Often, this can be done by making some reasonable assumptions about the distribution of θ in the population, or by collecting data and empirically estimating this function. Because the statistician can never be sure that the particular choice of a prior distribution is accurate, one criticism of Bayesian statistics is that one cannot be sure how well the posterior distribution represents the distribution of θ given x . The impact of the prior distribution on the final results of the analysis (i.e., the posterior

distribution) will vary depending on the statistician's choice for the distribution. Prior distributions that have significant influence on the posterior distribution are referred to as relatively "informative," whereas prior distributions with relatively little influence are called "non-informative" priors. For this reason, analysts tend to choose relatively non-informative prior distributions in situations where they are less confident about the distribution of θ in the population.

—William P. Skorupski

See also Bayesian Statistics; Posterior Distribution

Further Reading

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

PROBABILITY SAMPLING

Probability sampling is the term used to describe sampling where the probability of a population unit being selected in the sample is known. In nonprobability sampling, this probability is not known. When probability sampling is used to survey a population, population parameters can be inferred from the sample. Examples of probability sampling are simple random sampling, stratified sampling, and cluster sampling. In all of these examples, the probability that an individual unit is selected in the sample can be calculated.

For simple random sampling (without replacement), the probability that a unit appears in the sample is the same for every unit and is n/N . Consider a population of size five, say, five children. A simple random sample of size three is selected. This could be done by writing the children's names on five separate balls and putting the balls in an urn. The balls are mixed well and three balls are drawn out. The balls are all equal in size, shape, and weight, and so each ball will have an equal chance of being selected. The number of possible samples that could be selected is

$$\binom{N}{n} = \binom{5}{3} = 10.$$

The probability that an individual unit i is selected in the sample is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{6}{10} = \frac{3}{5},$$

where

π_i = number of samples that include the unit,

i = number of possible samples.

The numerator in this equation uses $N - 1$ and $n - 1$. If the sample includes unit i , then there are $N - 1$ units left in the population from which to choose the remaining sample of size $n - 1$. In the sample of children, after choosing the first ball out of five, there are only four balls left to choose the second ball and three balls left to choose the third ball in the sample. With probability sampling, sampling theory then can be used to estimate the sample mean, and the variance of the sample mean to infer the population mean.

Nonprobability sampling is called *haphazard sampling*, *convenience sampling*, and *judgment sampling*, among other terms. Examples of nonprobability sampling would be to stand on a street corner and interview pedestrians who walk by when the population of interest is all citizens who live in the city, running survey lines in a forest adjacent to roads when the population of interest is all the forest, and surveying only those businesses that appear to be representative. The samples could, in fact, give informative and reliable results. Pedestrians who walk may be representative of all the citizens, the forest adjacent to the road may be representative of the entire forest, and the judgment of what are representative businesses may be correct. However, without using probability sampling, sampling theory cannot be used, and there is no way of knowing if the resultant samples are informative and reliable. It is not possible to estimate how accurate the samples are or how precise they are.

—Jennifer Ann Brown

Further Reading

- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

PROFILE ANALYSIS

In psychological research, there are many instances in which the goal is to classify people as belonging to certain groups. Personality researchers often search for evidence of personality types, some intelligence researchers search for evidence that people exhibit different types of intelligences, and so on. The term *profile analysis* is used to represent a variety of techniques that share the common goal of empirically classifying individual observations into distinguishable groups based on common characteristics measured by observed variables. Strictly speaking, it is not a statistical technique because no inferences are being made to population parameters. Rather, profile analysis is a data reduction technique. It is conceptually similar to factor analysis, but the difference is that the focus in profile analysis is on grouping people (or observations) on the basis of common traits measured by observed variables.

The two major approaches to profile analysis are quite different from one another conceptually, but they share the same goal of grouping cases based on observed variables. The first approach is based on correlation methods, such as Pearson's r or Spearman's ρ . The goal under this approach is to group cases together that show similar patterns of spikes and dips across variables (i.e., the same shape), regardless of the absolute level of the scores (see Figure 1). Under the correlational approach, Persons A and B would be classified as belonging to one group (or profile) and Persons C and D to another.

By contrast, a second approach to profiling participants is based on measures of distance, such as Euclidean distance or Mahalanobis's distance (see Figure 1). Here, the emphasis is on creating groups based on the extent to which case scores are close in absolute value (i.e., level), regardless of the similarity

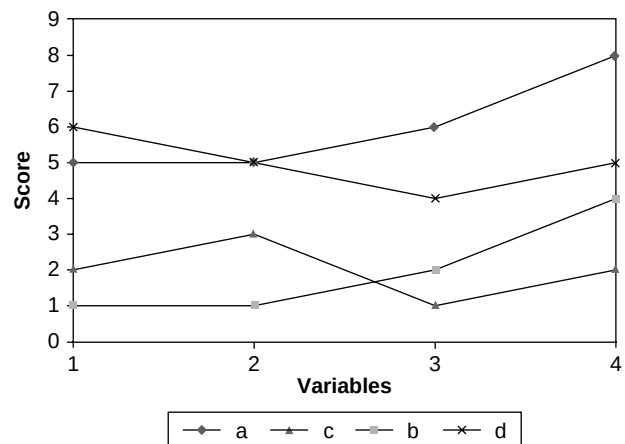


Figure 1 Examples of Different Score Profiles

of pattern shape. Under the distance approach, Persons B and C would be classified as belonging to one group and Persons A and D another.

Key Assumptions

There are two key assumptions of any profile analysis:

1. The sample is representative of the population.
2. There is a minimum of multicollinearity in the data. The reason that this is problematic is that variables that are collinear receive more weight in the solution.

The Case Study and the Data

Consider a practical example in which students were administered an experimental Advanced Placement test in Psychology. The test consisted of items primarily tapping one of four cognitive processing skills: memory processing, analytical processing, creative processing, or practical processing. The researchers were interested in examining the extent to which students showed different profiles of strengths and weaknesses across these four processing skills. Consequently, a Q-factor analysis (a correlation-based profile analysis) was performed on the students who took the test to determine the extent to which different profiles of achievement were observed. Using a principal component analysis (of the cases, not the variables), the researchers arrived at a solution yielding

six general profiles in the data. Overall, the researchers were able to group each of the 1,262 cases in the data sets as belonging to one of the six empirically distinguishable profiles.

—Steven Stemler

Further Reading

- Ding, C. S. (2001). Profile analysis: Multidimensional scaling approach. *Practical Assessment, Research & Evaluation*, 7(16). Retrieved October 1, 2005, from <http://PAREonline.net/getvn.asp?v=7&n=16>
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed., pp. 469–519). Upper Saddle River, NJ: Prentice Hall.
- Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (in press). Using the theory of successful intelligence as a basis for augmenting exams in AP Psychology and Statistics. *Contemporary Educational Psychology*.

PROJECTIVE HAND TEST

The Hand Test is a projective test consisting of a set of 10 cards approximately 3 inches by 5 inches in size. The first nine cards portray a hand in an ambiguous position, and the tenth card is blank. The cards are presented one at a time with the question, “What might this hand be doing?” The tenth card is blank and is given to the subject with the instructions: “This card is blank. I would like you to imagine a hand, and tell me what it might be doing?” Subjects are not limited in the number of responses they give to any individual cards or the entire set; however, subjects are encouraged with the instruction “anything else?” if they give only one response to the first card. The Hand Test elicits responses that reflect behavioral tendencies. Specifically, the Hand Test measures reactions that are close to the surface of personality, and reactions that are likely to be expressed in overt behavior. The administration time is typically brief (approximately 10 minutes) and the Hand Test is meant to supplement other material in a test battery. The Hand Test can be used with anyone old enough to verbalize a response (age 5+ years).

To score the Hand Test, the clinician simply classifies the responses according to clear-cut quantitative

scores (such as Aggression, Exhibition, Communication, Dependence, Acquisition, Tension, and Withdrawal), which reflect the person’s behavioral action tendencies in terms of how they interact with others and the environment. The qualitative scores generally reflect feelings and motivations underlying the expressed action tendencies. The Hand Test also consists of six summary scales, such as an index of overall pathology and an acting-out ratio, which is used to predict aggressive behavior.

Since its development in 1962, the Hand Test has been administered to more than a million people. The Hand Test manual was revised in 1983 and provides means, cutoff scores, and typical score ranges for normal adults and a number of diagnostic/clinical groups such as schizophrenia, mental retardation, organic brain syndrome, anxiety disorders, affective disorders, somatoform disorders, older adults, and personality disorders. A manual supplement provides norms for 5- to 18-year-olds and offers guidelines for interpreting child and adolescent responses and for integrating the Hand Test into a standard psychoeducational evaluation. Extensive research (more than 40 years’ worth) and independent reviews comment favorably on the Hand Test’s reliability and validity.

—Paul E. Panek

Further Reading

- Sivec, H. J., Waehler, C. A., & Panek, P. E. (2004). The Hand Test: Assessing prototypical attitudes and action tendencies. In M. Herson (Ed.), *Comprehensive handbook of psychological assessment* (Vol. 2, pp. 405–420). New York: Wiley.
- Wagner, E. E. (1983). *The Hand Test: Manual for administration, scoring, and interpretation*. Los Angeles: Western Psychological Services. (Originally published 1962)
- Young, G. R., & Wagner, E. E. (Eds.). (1999). *The Hand Test: Advances in applications and research*. Malabar, FL: Kreiger.

PROJECTIVE TESTING

Testing is a form of precise measurement. Some tests are merely exact observations. Others involve

manipulation of the environment and a cataloging of results or reactions. Projective testing goes one step further: Stimuli are presented, reactions are monitored and recorded, and then aspects of personality are deduced from these responses. The projection involved here is actually two-fold: It is used to refer, by some, to the deductive process of the examiner, and it also is used to hypothesize the process by which the respondent is assumed to have arrived at the response by projecting material of which he or she is unaware.

There are areas of behavioral testing where self-report and observations are heavily relied upon by the diagnostician. However, the assessment of personality, psychopathology, and character disorder does not readily lend itself to observation at an evaluation. At the same time, reports of symptoms (by significant others, or even by trained professionals) are often too inexact and complicated to be reliable. Self-report by the client himself or herself is subject to distortion, defensiveness, and misinterpretation. This is the realm of projective testing.

The logic of projective testing has several foundations. Its logic is inferential, whereas its technique is psychodynamic. In practice, the process of assessment relies heavily on the unconscious and, to some extent, on defense mechanisms. Of the latter, the defense of projection is most crucial, as one might expect from the name of the testing approach. The elements of this defense mechanism entail the tendency for a person to find unacceptable aspects of his or her personality manifestations elsewhere. Dynamically, the person is consciously unaware of these aspects, but is unconsciously troubled by them. The ability to maintain consistent repression of these threatening aspects is enhanced by projecting them onto other persons or situations.

Reality and psychological functioning levels often dictate the possible breadth and parameters of projection. Although a nonpsychotic individual who is unconsciously tormented by repressed aggressive impulses is not likely to see the devil on the bus, he or she may interpret comments as more aggressive than they actually were. However, the more ambiguous or undefined a situation is, the greater the possibility for it to become the object of projection.

Projective tests differ in the degree of ambiguity versus definitiveness of their stimuli. These aspects circumscribe the realm of these tests respectively. Below is a presentation of the major contemporary projective tests, each representing different points on this continuum.

The Rorschach is perhaps the best known test in this realm. It consists of 10 (horizontally symmetrical) inkblots featuring variations in color, shading, and forms. It represents a fairly nondefined set of stimuli, where respondents can hardly be “defying reality” regardless of what they reportedly perceive. As such, it enhances and encourages projection to a significant degree.

Contrary to common assumptions, the content of Rorschach responses is barely relevant in interpretation. Instead, there exists a complex scoring system of many composites and ratios, featuring such elements as location choice, form accuracy, three-dimensionality, color categories, shading, texture, organization of disparate parts, symmetry, and others. The significance of these elements was originally spelled out by armchair analysis, essentially reflecting experts’ opinions of the assumed projective origin of the elements. This is no longer the case. Empirical correlational research has revolutionized Rorschach methodology, so that the interpretations of elements and ratios are based almost totally on meticulously collected data. Today, Rorschach enjoys high scoring reliability and strong interpretive validity. Indeed, it is reasonable to question whether the Rorschach should still be classified as a projective test, because most of its current interpretive base is clearly operationalized and does not require any projective theory to formulate findings. It is noteworthy that Rorschach interpretations have a wide scope: from cognitive processes, to reality testing, to emotional control, to psychiatric diagnoses, and beyond. Furthermore, face validity of the test is not high. This implies that typical respondents cannot easily intuit what the test elements are intended to measure. As such, it is less susceptible to deliberate distortion or “faking” by respondents.

The Thematic Apperception Test (TAT) is another classic projective test that has evolved well beyond its

original intent. It consists of a collection of cards and drawings, most of which contain people in various forms of interactions or emotional contexts. First designed to elicit a hierarchy of psychological needs, this test has been usurped by psychodynamic diagnosticians of all types, with the assumption that the attitudes, feelings, and behaviors that a respondent attributes to the people in the pictures represent projections of his or her own dynamics.

The TAT has more limited response options than the Rorschach. First, the forms of its stimuli are actual representations, not inkblots. Second, the administration asks respondents to tell a story about the card, again a more constricting task than the Rorschach direction, which merely asks, What might this be? As a result, there is more of an issue of “reality constraints” for this test, but, at the same time, there will be much richer direct emotionality expressed in relation to the figures as opposed to those expressed to inkblots. In addition, the scope of the TAT is more oriented to interpersonal relationships and self-reflection, and is hardly appropriate to elicit a clear picture of cognitive organization the way the Rorschach does.

The rather straightforward interpretation of projected attitudes or emotions as representing those of the respondents’ lends face validity to TAT test interpretation. Ironically, it is probably this face validity that has shielded the TAT from the kind of sophisticated psychometric development the Rorschach has witnessed. To date, there is no rigorous or reliable coding or interpretive method for the TAT. Instead, there are a host of interpretive schemes that are part of a larger, poorly operationalized methodology of fantasy interpretation. This renders the precision of the TAT interpretation not much more scientific than that of dream interpretation. This lack of a precise coding strategy for this test has also hampered empirical validity research for this test. However, some studies have been published using somewhat reliable methodologies, and these have made their way into the diagnostic literature. Nonetheless, it remains a popular diagnostic tool and is featured widely in interpretive reports because professionals of various disciplines who are not psychometrically trained can easily relate to its constructs.

Projective figure drawings come in a wide array of tests such as Draw a Person, the House-Tree-Person Test, the Family Kinetic Drawing Test, and others. The projective hypothesis here is that all of these drawings are self-representations. The literature here is not consistently systematized, but there are major authors whose work is authoritative across tests. Presently, there are hosts of systems and schemes that interpret various aspects of drawings. For example, small human drawings are often taken as evidence of projected low self-esteem, bare branches as signs of inefficacy, and the omission of a house entrance as evidence of defensiveness.

Two major limitations plague figure drawings. First, empirical data, other than compilations of anecdotal reports, are few. Validation studies that are well designed do not offer empirical support for the test. Another limitation is the confounding effects that poor eye-hand (especially graphomotor) coordination entails in drawings. Distortions that are interpreted projectively may, in fact, be due to intrinsic inability of examinees to draw properly regardless of what they are drawing.

Nonetheless, figure drawings are very popular as projective tests. It is noteworthy that they are progressively more popular in the literature of fields allied with psychology (e.g., art therapy) than they are in diagnostic psychology proper.

Incomplete sentence tests mark another point on the projective testing continuum. The best known exemplar of this test (by Julian Rotter, a social learning theorist) was, in fact, not designed as a projective. It features different 40-item sets of sentences that begin with several words (e.g., “Back home . . .” or “A mother . . .”), each followed by a blank that the respondent fills in. There are separate sets of items for different developmental/age ranges.

It should be noted that, other than the basic coding systems, which were originally designed by Rotter (and other technical variations), there is no standard projective scoring and interpretive system for this instrument. Clinicians who use it as a projective test often use similar (unstandardized) interpretive strategies here as they do for the TAT.

There are some psychometricians who question the propriety of classifying the Incomplete Sentences Blank as a test altogether. These critics see the instrument as a standard questionnaire that elicits information about one's attitudes and feelings, and assert that it might best be described as a typical intake tool one would use at a clinic instead of, or in conjunction with, a background interview. There is some evidence, nonetheless, that some respondents find the task of completing sentences distracting enough to lower their defenses. The result is that more information is revealed than one might expect to discover in a regular interview structured by a question-and-answer format.

Still, labeling the sentence completion task as a psychological test is equivocal, at best. Moreover, classifying the instrument as a projective test is seen by some as more incongruous yet. (It should be noted that some diagnosticians emphasize ancillary features, such as the significance of omitted responses, changed responses, or inordinate deliberations/hesitations for specific items, as the more projective anchors of this test.)

The critiques that challenge the status of this "test" are most salient, because the standard printed instructions for sentence completion tests (i.e., Rotter's) typically ask the respondent to answer truthfully in a manner that reflects his or her true feelings and opinions. Some clinicians sidestep this issue by deleting the instructions and then presenting alternate instructions. These alternate instructions present the sentence completion task as an academic exercise designed to test the respondent's approach to writing and sentence construction. The projective hypothesis is then appealed to, with the assumption that respondents will use the opportunity to project their inner conflicts and unacceptable emotions into this allegedly neutral task.

Although this alternative approach can be challenged as deceptive and as ethically questionable, many of the other projective tests cross that ethical line to some extent. Consider, for example, the standard instructions for the Rorschach, which state that "there are no right or wrong answers." In fact, there are precise statistical frequency tables that the examiner uses in interpreting the reality testing of the

respondent, belying the denial of right and wrong. Similarly, the TAT is often presented to respondents in a less-than-transparent approach, usually as a test of the creativity of the respondent in "making up a story." Some of the popular instructions include the examiner telling the respondent a script, which is approximately as follows: "You will be the author, and I will be the secretary. You can make up any story. There are no right or wrong answers here. All that matters is your imagination." Although it is true that there are no formal tables for the TAT, which resembles the Rorschach Form Quality tables, there are, in fact, accepted standards of what kinds of stories are appropriate or inappropriate for each card. Despite the fact that the appropriateness here is more on the emotional level and less related to reality testing, some deception seems to be present here. It is not surprising, however, that many diagnosticians are partial to projectives, and these specialists justify the misinformation as a necessary aspect of testing that puts respondents off guard to allow access to dynamics that are typically inaccessible and hidden—even from the respondent.

Projective testing—from the mystery-shrouded and infamous Rorschach, to the more intuitive TAT, to the somewhat obscure figure drawings, to straightforward sentence completion—was prominent in the beginning of the discipline of psychological testing, when psychometrics were neither systemized nor operationally defined. As more objective tests, systematized observations, and behavioral measures came to dominate diagnostics, projectives were marginalized in many professional circles and relegated, at best, to ancillary roles. (This pattern was least evident in specific urban centers in the United States, such as New York City, Boston, and San Francisco, where psychoanalysis still reigns. In addition, countries from the former Soviet Union, where Freudian thinking was banned for decades, have seen a keen interest in psychoanalysis as a theory and projectives as a methodology.) In response to the supremacy of objective diagnostics based on the *Diagnostic and Statistical Manual of Mental Disorders*, the past two decades have witnessed a resurgence of projective testing as a psychometrically

kosher approach, a move championed by operationally definitive Rorschach studies accompanied by impressive clinical validation research. Although stereotypical dismissive attitudes still prevail in many psychological circles (particularly in the Midwest bastions of behaviorism, where psychoanalysis, projective testing, and astrology are grouped together as “pseudo-sciences”), projective testing is now a common feature in clinical diagnostic, which is used in tandem with more objective psychological measures to yield a comprehensive picture of personality and pathology.

—Samuel Juni

Further Reading

- Groth-Marnat, G. (2003). *Handbook of psychological assessment*. New York: Wiley.
- Juni, S. (1979). Theoretical foundations of projection as a defence mechanism. *International Review of Psychoanalysis*, 6, 115–130.
- Rapaport, D., Gill, M., & Schaffer, R. (1968). *Diagnostic psychological testing*. New York: International Universities Press.
- Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing*. New York: Grune & Stratton.

PROPENSITY SCORES

A propensity score is the conditional probability of a unit being assigned to a condition given a set of observed covariates. These scores can then be used to equate groups on those covariates using matching, blocking, or ANCOVA. In theory, propensity score adjustments should reduce the bias created by nonrandom assignment, and the adjusted treatment effects should be closer to those effects from a randomized experiment.

In a randomized experiment, a unit's true propensity score is the known probability of being assigned to either the treatment or comparison condition. For instance, when using a coin toss, the true propensity score would be $P = .5$ for every unit. In nonrandomized studies, we do not know the true probability of

being in a condition; therefore, it must be estimated using observable variables.

Computing Propensity Scores

Propensity scores are probabilities that range from 0 to 1, where scores above .5 predict being in one condition (i.e., the treatment group) and those below .5 predict being in the other condition (i.e., the comparison group). Traditionally, propensity scores are computed using a logistic regression; however, more recently, other methods have been used, such as classification trees, bagging, and boosted modeling.

Logistic regression is the most common method for computing propensity scores. In this method, a set of known covariates is used in a logistic regression to predict the condition of assignment (treatment or control), and the propensity scores are the resulting predicted probabilities for each unit. The model for this regression equation is typically based on variables that affect selection into either the treatment or the outcome, including interactions among the predictors. It is not necessary that all predictors be statistically significant at $p < .05$ to be included in the model.

Classification and regression trees predict categorical outcomes (often dichotomous) from predictor variables through a sequence of hierarchical, binary splits. Each split is determined by the predicted probability that a unit will select into conditions based on a single predictor. With each dichotomous split, two branches result, and the splitting process continues for each new predictor until a certain number of nodes is obtained or all significant predictor variables are used. The result is a binary tree with terminal nodes (branches) representing groups of units that have the same predicted condition, although each node may have reached the same condition using different predictors. The predicted outcomes are propensity scores. A disadvantage of classification and regression trees is that their results are not very robust. The modeled trees are highly variable, and splits often change with minor variations in the data. One way to increase the stability of trees is to use bagging.

Bagging (bootstrap aggregation) averages results of many classification trees that are based on a series

of bootstrap samples. In this case, random samples (with replacement) are drawn from the observed data set and additional observations are simulated to mimic the observed distributions. A new classification tree is computed for each simulated data set. Bootstrapped trees are aggregated to form aggregated trees, resulting in a more stable prediction model.

Boosted modeling, like bagging, uses an algorithm to create multiple models using different predictors. Each model is weighted based on the strength of the model and aggregated to form a single, more stable model. It differs from bagging by (a) using the full, original sample for each model, whereas bagging uses bootstrap samples; and (b) weighting units based on how difficult they are to classify. Each iteration assesses how well each unit was correctly classified, and then assigns a greater weight to that unit for the next iteration. Boosted modeling is used with either logistic regression procedures or classification trees.

Regardless of the method used to create propensity scores, the primary goal is not to optimally predict assignment to conditions, but to balance the distributions of covariates over conditions. This is typically assessed by using a factorial ANOVA to test the interaction between the propensity score strata and conditions (as independent variables) on the covariates. A nonsignificant interaction indicates that the distributions are balanced.

Predicting Assignment for More Than Two Conditions

When propensity scores are to be estimated for units selecting into more than two conditions, devising these estimates is more complicated. Three methods for doing so are as follows:

1. *Paired comparisons*, in which a separate propensity score model is created for each possible pair of conditions (i.e., three conditions have three possible pairs and will have three different propensity score models)
2. *Ordinal predictions*, in which an ordinal logistic regression is used to predict ordered dose-like levels of conditions (i.e., no treatment, partial treatment, full treatment)

3. *Multivalued treatments using a generalized propensity score*, which computes the probability of receiving a certain dose or level of the treatment, but can be used for nominal or ordinal categories using multinomial or nested logit models.

Adjustments Using Propensity Scores

Once propensity scores are calculated, statistical adjustments can be made using those scores. Typically, this is done using matching, stratification, covariate adjustment, or weighting.

Matching identifies similar cases from experimental and control groups based on the proximity of their propensity scores. Methods for matching include exact matching (cases paired on exactly the same scores), and caliper, optimal, or greedy matching, in which scores are paired by proximity. The most preferred methods of matching are greedy matching, which selects the closest matches first and then progressively farther matches, and optimal matching, which finds the smallest cumulative distance from all possible pairs. To prevent dropping treated participants from the analyses, some researchers recommend using index matching, which matches treated participants with multiple control participants. Others recommend matching proportionately more participants in the control group to fewer participants in the treatment group.

Stratification divides the entire distribution into four to seven strata based on propensity scores, so observed variables are balanced for treated and control units within each stratum. To estimate the adjusted treatment effects, treatment and control group means are computed as the unweighted average of the cell means over strata for each group. Although stratification usually permits the inclusion of all cases, some cells may contain few or no units. Reducing the number of strata or changing the cut-points that define the blocks may reduce this problem, but determining ideal cut-points for strata can be difficult. Despite these concerns, stratification is often a favored method because it does not require modeling nonlinear trends, like covariate adjustments and weighting, and is less complicated than matching.

Covariate adjustment uses propensity scores as covariates in an analysis of covariance in order to

remove bias due to the covariates from the effect estimate. Although some studies have indicated that covariate adjustment can be an effective method for removing bias, Paul R. Rosenbaum and Donald B. Rubin do not recommend this method, particularly when variances among variables that created the propensity scores are heterogeneous.

Weighting attempts to balance treatment and control groups by multiplying (weighting) observations by the inverse of the propensity score. Weighting often uses propensity scores that are computed using nonparametric statistics, such as a series estimator, or semiparametric statistics, such as binary regression quantiles, as well as parametric statistics, such as logistic regression. The adjusted treatment effect, $\hat{\beta}$, is estimated using a formula, such as this one:

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^N \frac{Y_i(T_i)}{e(x_i)},$$

in which Y_i is the outcome variable, T_i is the treatment condition, and $e(x_i)$ is the propensity score for each unit.

Conclusions

Although the effectiveness of propensity scores is still under scrutiny, some work suggests that when propensity scores are used in accordance with the guidelines described by Rosenbaum and Rubin, they can reduce bias in quasi-experiments. However, even when used properly, propensity scores cannot adjust for hidden bias, which occurs when relevant covariates are not included in the propensity score model. Propensity scores can be only estimated from observed variables; however, many factors that are not measured may contribute to selection bias.

—M. H. Clark and William R. Shadish

Further Reading

- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to

data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2, 259–278.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–710.

Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology*, 150(4), 327–333.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516–524.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249–264.

Shadish, W. R., & Clark, M. H. (2002). An introduction to propensity scores. *Metodologia de las Ciencias del Comportamiento Journal*, 4(2), 291–298.

StatSoft, Inc. (2003). *Electronic textbook*. Retrieved June 14, 2005, from <http://www.statsoftinc.com/textbook/stathome.html>

PSYCHOLOGICAL ABSTRACTS

Psychological Abstracts (PA, 1927–present) is a monthly publication produced by the American Psychological Association (APA). The publication provides bibliographic citations and abstracts of scholarly literature pertaining to the field of mental health and behavioral sciences. PA was preceded by the *Psychological Index* (1894–1935). The electronic version of the *Psychological Abstracts* is the PsycINFO database, first published in 1967. Literature coverage in PsycINFO spans from 1887 to present for more than 50 countries and more than 25 languages. It includes journal articles; book chapters and books; dissertations; technical reports; and other publications such as conference proceedings, program evaluations, or psychological assessment instruments. The database is updated weekly and contains more than 2 million publication records.

Each monthly issue of PA consists of the following parts: (a) an author index, (b) a brief subject index

based on controlled vocabulary, (c) a thesaurus of psychological terms used in the subject index, (d) a book title index, and (e) abstract volumes that present citations and nonevaluative summaries. An expanded and cumulative author and subject index is published annually and placed at the end of each volume. Searches on the PA can be conducted by author or subject by referring to the author and brief subject indexes in the back of each monthly issue. The main concepts of each cited document are summarized in up to five major terms or descriptors printed in the brief subject index. The subject terms used in this index are based on controlled vocabulary and scientific terminology provided in the *Thesaurus of Psychological Index Terms*.

Each PA record is divided into several fields. For instance, book entries contain the following elements in order: (a) record number, (b) author(s) or editor(s), (c) affiliation of first author, (d) book title, (e) publisher and bibliographic information, (f) table of contents listing PA record number for each chapter, and (g) quoted material indicative of the content. Abstracts are provided only for journal article entries. Because PA uses the same abstract numbers each year, abstract numbers should be looked up in the same year/volume as the index used.

PA is available through individual- or institution-based subscriptions obtained from the APA. Most academic institutions provide access to both PA and PsycINFO to their faculty, staff, and students. PA does not provide full text of the documented records. Full-text articles can be obtained through other APA services such as PsycARTICLES or directly from libraries or through interlibrary loan services.

—Marjan Ghahramanlou-Holloway

Further Reading

- Reed, J. G., & Baxter, P. M. (1992). Psychological Abstracts and PsycBOOKS. In J. G. Reed & P. M. Baxter, *Library use: A handbook for psychology* (2nd ed., pp. 37–56). Washington, DC: American Psychological Association.
- Williams, R. (1982). Comprehensiveness of Psychological Abstracts. *American Psychologist*, 37(3), 339.

PsycINFO Psychological Abstracts: <http://www.apa.org/psycinfo/>

Thesaurus of Psychological Index Terms, 10th edition: <http://www.apa.org/books/3100083.html>

PSYCHOMETRICS

Psychometrics is the systematic application of mathematical principles in order to measure (quantify) psychological constructs. Measurement consists of rules for assigning symbols to objects to (a) represent quantities of attributes numerically (scaling), or (b) define whether the objects fall in the same or different categories with respect to a given attribute (classification). The objects of study in psychometrics are usually people, but many studies involve other animals, or basic biological or psychological processes. Accurate measurement is essential for the development of any science because it permits objective quantification and communication of findings. Standardized measurement specifies the concept in which we are interested so that others can replicate or refute our work. Psychometric theory has been applied in a wide range of fields, including the measurement of intelligence, personality, attitudes and beliefs, and achievement, as well as in health-related disciplines.

Primary Approaches

Psychometric theory has evolved into several areas of study. Early psychometric work was in the area of intelligence testing. This work gave rise to classical test theory (CTT). More recently, item response theory has emerged to address some of the shortcomings of the CTT approaches.

Classical Test Theory

Core to the CTT approach are the concepts of reliability and validity. Reliability is the study of consistency in measurements. All measures contain some degree of error or imprecision, and CTT procedures are used to evaluate the amount of error present in a

measure and whether the observed error is systematic or random. Reliability is assessed statistically, often with correlational techniques. For example, test-retest reliability (stability) of a test or the reliability of alternate forms of a test is evaluated with the Pearson coefficient.

Validity, which is dependent on a measure first being reliable, refers to whether a measure is measuring what it is supposed to measure. Validity often is a multifaceted and ongoing research process that involves establishing various aspects of a measure's validity. For example, concurrent validity refers to a measure's correlation with a known criterion measure collected at the same time, whereas predictive validity is the ability of a measure to predict a criterion at some future point in time.

Item Response Theory

Item response theory (IRT) models the relationship between latent traits and responses to test items. In contrast to CTT approaches, IRT methods have the potential to produce measures that are (a) sample independent—item parameters can be created independent of the sample taking a test; (b) falsifiable models—direct tests are available to evaluate whether a specific IRT model is appropriate for a particular data set; and (c) directly addressing measurement bias at the individual item level before these items are combined to form a scale score.

Statistical Methods

The field of psychometrics uses a number of statistical procedures, particularly correlation and regression techniques. Multivariate descriptive methods frequently are used and include factor analysis to uncover underlying dimensions in a test data set, cluster analyses to find test items or test respondents similar to each other, and multidimensional scaling to find simple representations for complex test data. More complex nonlinear regression models are used to test IRT models.

—Thomas E. Rudy

Further Reading

- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

PsycINFO

PsycINFO, produced and copyrighted by the American Psychological Association (APA), is an electronic database of bibliographic citations and abstracts of scholarly literature pertaining to the field of mental health and behavioral sciences, and psychology-related aspects of other disciplines. The literature coverage in PsycINFO, spanning from 1887 to present, covers more than 50 countries and more than 25 languages. It includes journal articles; book chapters and books; dissertations; technical reports; and other publications such as conference proceedings, program evaluations, and psychological assessment instruments. The database is updated weekly and contains more than 2 million publication records. Since its first publication in 1967, PsycINFO has become one of the most widely used resources for scholarly and professional research.

The PsycINFO database is an electronic version of the print publication *Psychological Abstracts*. As of August 2005, PsycINFO's journal coverage accounts for 78% of the database and includes 1,980 titles, 98% of which are peer reviewed. The journal coverage list changes monthly as journals are added or discontinued. Book chapters from edited books make up 7% of the database; authored and edited books make up 3%. The final 12% of the database is devoted to dissertations and other secondary publications. Nearly all publication records, from 1995 to present, provide abstracts and are provided in APA-style format.

Literature searches on PsycINFO records can be conducted in various ways. Records are divided into fields. The content of each field varies depending on

the type of literature documented within the database. For instance, one can complete a search by using one or more of the following fields: author, title, keywords, document type, language, population group, or publication year, or by choosing additional fields not mentioned here. Search results are listed by record date, beginning with the most recently published. Each record, at a minimum, generally contains the name and affiliation of the author, title, source, abstract, and publication year. To maximize the precision of research efforts, PsycINFO users can refer to the *Thesaurus of Psychological Index Terms*, which provides the most up-to-date controlled vocabulary terms and scientific terminology.

Access to PsycINFO is available through individual- or institution-based subscriptions obtained from the APA. Most academic institutions provide access to PsycINFO for their faculty, staff, and students. PsycINFO does not provide access to the full text of the documented records. Full-text articles can be obtained through other APA services such as PsycARTICLES or directly from libraries or through interlibrary loan services. Publishers and/or authors interested in submitting their publication for coverage in PsycINFO may do so by following the APA guidelines provided at <http://www.apa.org/psycinfo/about/covinfo.html>.

—Kathryn Lou and
Marjan Ghahramanlou-Holloway

Further Reading

- American Psychological Association. (1992). *PsycINFO user manual*. Washington, DC: Author.
- Gosling, C. G., Knight, N. H., & McKenney, L. S. (1990). *Search PsycINFO: Instructor guide*. Washington, DC: American Psychological Association.
- PsycINFO Psychological Abstracts: <http://www.apa.org/psycinfo/>
- Thesaurus of Psychological Index Terms*, 10th Edition: <http://www.apa.org/books/3100083.html>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Lawrenz, F., Keiser, N., & Lavoie, B. (2003). Evaluative site visits: A methodological review. *American Journal of Evaluation*, 24(3), 341–352.

Site visits are a commonly employed, but little discussed, evaluation procedure. The purpose of this paper is to review the state of the art regarding site visits, as well as to catalyze a discussion of site visits centering on the question of whether or not existing practices constitute a set of methods or a methodology. The researchers define evaluative site visits and review the current state of practice (using **PsycINFO**) in site visits. They also outline the differences between methods and methodology and consider whether or not the current state of practice constitutes a methodology.

ENCYCLOPEDIA OF
MEASUREMENT

AND
STATISTICS

ENCYCLOPEDIA OF
MEASUREMENT
AND
STATISTICS

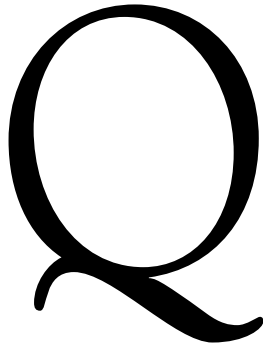
VOLUME **3**

EDITOR
NEIL J. SALKIND
UNIVERSITY OF KANSAS

MANAGING EDITOR
KRISTIN RASMUSSEN
UNIVERSITY OF KANSAS

A SAGE Reference Publication

 **SAGE Publications**
Thousand Oaks ■ London ■ New Delhi



If the Aborigine drafted an I.Q. test, all of Western civilization would presumably flunk it.

—Stanley Garn

Q METHODOLOGY

Q methodology was invented by British psychologist and physicist William Stephenson (1902–1989) and provides both the tools and the philosophy for the systematic study of subjective behavior. Stephenson first introduced his innovation in 1935; however, it was publication of his *The Study of Behavior*, in 1953, that brought his ideas to a wider audience, and it is in this book that he fully explores the utility of Q methodology for researchers interested in studying subjectivity.

Elements of a Q Approach

Concourse and Q Samples

The concept of *concourse* refers to the stream of commentary revolving around any topic, be it the price of gasoline, the war in Iraq, the steroid scandal in baseball, or any other topic. These comments are matters of opinion and are shared with others in the culture. Researchers can gather concourses from a number of sources (e.g., interviews, texts,

etc.), with interview responses having the virtue of being in the natural language of participants. From this concourse, a selection of statements is drawn that constitutes the *Q sample*, which may be either structured or unstructured. Whereas an unstructured sample is composed unsystematically and is made up of statements relevant to the subject under consideration, a structured Q sample usually embodies a theoretical framework, with greater attention given to coverage of subissues in the statement selection process. Q samples are typically structured in terms of the principles of factorial experimental design, with statements provisionally placed into categories of the design. However, unlike the case in scale construction, there is no assumption that these categories will (or should) hold together once the participants begin to respond to these statements.

P Set

Individuals who serve as study participants are known as the *P set*, and their selection is a function of the research question. Participants may be chosen

because of theoretical concerns thought to be relevant. For example, a Q study concerning health care in the United States might predispose the researcher to try to get doctors, lawyers, patients, hospital administrators, politicians, insurance company representatives, and so on as participants. On the other hand, a Q study of understandings of the Judy Garland film *The Wizard of Oz* might lead the researcher to seek out participants who are readily available, assuming almost everyone has a level of familiarity with that classic film.

P sets in Q studies are typically small, most often 40 or fewer. The aim is to allow subjective viewpoints to be revealed, not to make the claim that all possible viewpoints are represented or that a certain percentage of the population holds a particular view.

Q Sorting

Participants are given the Q sample (each statement being printed on an individually numbered card) to rank order along an established continuum according to a “condition of instruction,” typically “most agree” to “most disagree.” However, the condition of instruction may be “most like me” to “most unlike me” or “most like my view in 20 years” to “most unlike my view in 20 years” and so on. Conditions of instructions can be used to create experimental conditions for testing theories. For example, conditions of instruction might be constructed to tap into interpersonal dynamics, for example, “most like the view my father would like me to have” to “most unlike the view that my father would like me to have.”

It is important for participants to have enough workspace so that all statements can be arrayed in front of them. Participants should first read through the cards in order to gain familiarity with them and then begin to make piles of statements that they agree or disagree with or feel neutral toward. The participant then begins to identify the most agreeable statements from the “agree” pile, placing them under the most positive end of the continuum, for example, +4. The

Table 1 Scoring Form

Most Disagree *Most Agree*

-4 -3 -2 -1 0 +1 +2 +3 +4

(2) (2)

(3) (3)

(4) (4)

(5) (5)

(7)

sorter then singles out the next-most-agreeable statements and places them under +3. It is then advisable to turn to the “disagree” pile and identify those statements with which the person most disagrees (−4). This process is continued until each statement has found a place under the continuum. Sorters are then encouraged to examine the array and move statements from one column to another if they choose. Completed Q sorts are formal models of the sorter’s point of view with respect to the topic under study. Table 1 shows a possible scoring form, ranging from −4 to +4 for a Q sample of size $N = 35$ statements.

For theoretical purposes, Q sorters are encouraged to follow a quasi-normal distribution, requiring the person to think systematically about the statements and to make more distinctions. Some have criticized this “forced” distribution as creating an artificial constraint on the sorter, arguing for a “free” distribution in which the sorter can place any number of statements at any point along the continuum. The literature in Q methodology has consistently demonstrated the differences in forced-free distributions to be statistically insignificant; however, a forced distribution is preferred for the theoretical reasons stated above.

It is also advisable to conduct post-sort interviews with participants to gain deeper understandings of why they ordered the statements in the manner that they did. How did they construe the meanings of the statements that were of most salience (e.g., +4, -4),

and why were they determined to be of such importance? The statements placed in the more neutral columns (i.e., +1, 0, -1) may also be important for eliciting more information. Generally, these statements are placed in these middle positions because there is a feeling of neutrality (in a relative sense) toward these items. However, at times, statements are placed in these columns because sorters feel conflicted about them; that is, they may agree strongly with one part of the statement and disagree strongly with another part. Thus, a deeper understanding of why this statement was placed where it was will inform the researcher.

Statistical Procedures and Factor Interpretation

Once collected, the data are entered into dedicated statistical programs such as PQMETHOD or PCQ (both accessible at <http://www.qmethod.org>), and correlations are calculated between persons (not between items). The correlation matrix is then factor analyzed to reveal groups of individuals with shared viewpoints. In the more conventional R methodology, the correlating and factoring are performed on traits or scores on objective tests, whereas in Q, the correlating and factoring are done on the subjective assessments of persons. That is, contrary to widespread opinion, a Q analysis is not simply based on a transposed R data matrix; rather, R methodology focuses on matrices of objective scores (e.g., IQ scores, scores on authoritarianism scales), whereas Q focuses on entirely different matrices of subjective impressions.

By way of illustration, suppose that 10 individuals have sorted 35 statements concerning their views of the war in Iraq, resulting in the hypothetical factor structure in Table 2. Factor loadings in this case are statistical measures of the degree of correlation with that factor.

The data in Table 2 show that persons 1, 2, 4, 6, 7, 8, and 10 have statistically significant factor loadings on Factor A, while persons 3, 5, and 9 have statistically significant factor loadings on Factor B. The formula for determining the significance threshold for a factor loading at the .01 level is $2.58(1/\sqrt{35}) = .44$, where

Table 2 Factor Loadings in the Iraq War Study

	A	B
Person 1	.88	.21
Person 2	.76	.04
Person 3	.15	.81
Person 4	.59	.19
Person 5	.08	.89
Person 6	.71	-.19
Person 7	-.88	.15
Person 8	-.77	.20
Person 9	.12	.75
Person 10	-.70	.14

35 is the number of statements. Fundamentally, this means that persons with significant factor loadings on Factor A share a common viewpoint, while those with significant factor loadings on Factor B share a separate viewpoint. An array of factor scores (i.e., a composite Q sort) is estimated for each of the factors, and it is this array that provides the basis for interpretation.

Note that Factor A is bipolar; that is, that the significant loadings run in both positive and negative directions. This means that the statements judged “most agreeable” by persons at the positive end of the factor are the statements considered “most disagreeable” by those at the negative pole. In other words, Q sorters at both ends of the factor are responding to the same themes, but in reverse order. In the hypothetical study of the Iraq War, for instance, imagine that Factor A arises from partisanship, with those at the positive end of the factor (persons 1, 2, 4, and 6) being Republican supporters of President Bush’s war against terror and those at the negative end of the pole (persons 7, 8, and 10) being Democrats who just as avidly rejected the same statements that were so warmly embraced by the Republicans (and vice versa). The Q sorters who defined Factor A were thus seeing the statements through the prism of their own partisanship, and this would constitute the theme of Factor A.

Factor B is orthogonal to Factor A and thus represents a wholly different point of view. Once again, factor interpretations are reached through examining the array of statements as produced by Factor B. Assume hypothetically that the statements rank highest (+4, +3) had to do with concerns over preemptive war and the standing of the United States in the world community, and so on. Assume further that other statements receiving high positive scores did not hold President Bush responsible for the war and expressed the view that any U.S. president would have probably acted in the same manner. The theme of the factor would thus be concerned with larger, geopolitical issues that transcend partisan politics.

Q Methodology and the Single Case

Q methodology is also well suited for use in the study of single cases. What is of significance in Q methodology is the participant's point of view rendered manifest by the procedure of Q sorting; consequently, issues of reliability and validity (in an R-methodological sense) recede in importance. Single-case studies proceed in much the same manner as outlined above, the chief difference being that the individual performs a number of sorts under multiple conditions of instruction, with the factors pointing to different tendencies and response predispositions within the same individual.

—James C. Rhoads

See also Measurement; Questionnaires; Validity Theory

Further Reading

- Brown, S. R. (1971). The forced-free distinction in Q-technique. *Journal of Educational Measurement*, 8, 283–287.
- Brown, S. R. (1986). Q technique and method. In W. D. Berry & M. S. Lewis-Beck (Eds.), *New tools for social scientists*. Beverly Hills, CA: Sage.
- Rhoads, J. C. (2001). Researching authoritarian personality with Q methodology. Part I: Revisiting traditional analysis. *Operant Subjectivity*, 24(2), 68–85.
- Rhoads, J. C. (2001). Researching authoritarian personality with Q methodology. Part II: Intensive study. *Operant Subjectivity*, 24(2), 86–103.
- Stephenson, W. (1953). *The study of behavior*. Chicago: University of Chicago Press.

Stephenson, W. (1977). Factors as operant subjectivity. *Operant Subjectivity*, 1, 3–16.

Stephenson, W. (1982). Q-methodology, interbehavioral psychology, and quantum theory. *Psychological Record*, 32, 235–248.

Thomas, D. B., & McKeown, B. F. (1988). *Q methodology*. Beverly Hills, CA: Sage.

Q-Q Plot

The quantile-quantile, or Q-Q, plot is a graphical procedure used to visually assess goodness of fit to a particular distribution. While the Q-Q plot will not provide a numerical measure of the goodness of fit, unlike such common tests as the chi-square goodness-of-fit test, the Kolmogorov-Smirnov test, or the Cramer-von Mises test, the graph can provide insight into how a given data set deviates from the specified distribution.

There are two common situations in which a Q-Q plot is used. The first of these applications is in determining whether two data sets are from a common distribution. The following steps are used to construct the plot:

1. Order each of the two samples to obtain the n order statistics $x_{1:n}, x_{2:n}, \dots, x_{n:n}$ and $y_{1:n}, y_{2:n}, \dots, y_{n:n}$.
2. Plot the ordered pairs $(x_{i:n}, y_{i:n})$ and examine the plot for linearity. If the samples are identical, a line with slope 1 and intercept 0 is obtained. If the points come close to fitting such a line, it is concluded the two samples are drawn from a common distribution.

This use of the Q-Q plot is described in further detail online both at Eric Weisstein's MathWorld and in the National Institute of Standards and Technology *e-Handbook of Statistical Methods*.

The second common application of the Q-Q plot is in testing for univariate normality. This is often referred to as a *normal probability plot*. The following steps are used to construct a Q-Q plot when testing for normality:

1. Order the n original observations $x_1, x_2, \dots, x_n$ to obtain the n order statistics $x_{1:n}, x_{2:n}, \dots, x_{n:n}$.
2. Find the n normal quantiles $q_1, q_2, \dots, q_n$, where q_i is the standard normal quantile associated with the i th

order statistic. The formula for q_i is given by

$$q_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right), \text{ where } \frac{i - 3/8}{n + 1/4}$$

is a “continuity correction” or a “plotting position.”

Another common plotting position is $\frac{i - 1/2}{n}$.

3. When testing for fit to a distribution other than the normal distribution, the q_i 's are found by replacing the inverse normal function in Step 2 with the appropriate inverse function.
4. Plot the pairs $(q_i, x_{i:n})$ and examine the plot for linearity. If the data are univariate normal, the plot will be a straight line.

Example

Consider the data set of 15 numbers in Table 1. These numbers were generated from a normal distribution with mean 100 and standard deviation 15, so the Q-Q plot would be expected to have a linear pattern. The computations needed are shown in Table 1, using

Table 1 Computations for Q-Q Plot

$x_{:i}$	$p_{:i}$	$q_{:i}$
57.01	0.041	-1.739
70.14	0.107	-1.245
82.24	0.172	-0.946
90.49	0.238	-0.714
91.31	0.303	-0.515
92.62	0.369	-0.335
95.54	0.434	-0.165
96.61	0.500	0.000
97.19	0.566	0.165
100.45	0.631	0.335
101.92	0.697	0.515
103.16	0.762	0.714
106.90	0.828	0.946
109.03	0.893	1.245
123.35	0.959	1.739

Blom's plotting position. In Figure 1, the observed values x_i are on the vertical axis, and the theoretical quantiles q_i are on the horizontal axis. The choice of axes is arbitrary since the main feature of interest is the linearity of the plotted points. The Q-Q plot was created with the free statistical package R. Similar plots can be made with any standard statistical software package such as SAS or SPSS. Some handheld graphing calculators can even make normal probability plots.

While the points do not fall perfectly on a straight line, particularly in the lower-left and upper-right portions of the Q-Q plot (see Figure 1), this plot does not indicate a serious deviation from normality that might prevent an analyst from using a standard statistical technique that requires the normality assumption. The Q-Q plot will also identify outliers; they will be seen as points in the lower-left or upper-right portions of the plot that deviate from the general linear pattern formed by the remaining observations. The analyst can then examine the outliers and decide how to proceed with the statistical analysis.

Figure 2 demonstrates what a Q-Q plot would look like when the data are drawn from a heavily right-skewed distribution. In this case, a random sample size of 15 was drawn from the exponential distribution, with the mean equal to 1, and the Q-Q or normal probability plot was again constructed using R. Notice how this scatterplot is much more curved in the middle portion of the graph than in the previous example.

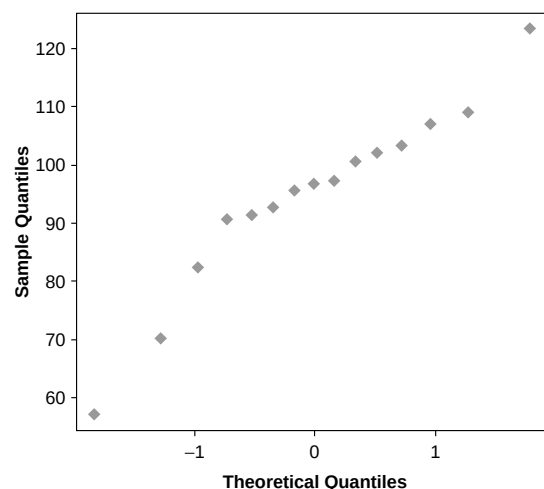


Figure 1 Normal Q-Q Plot

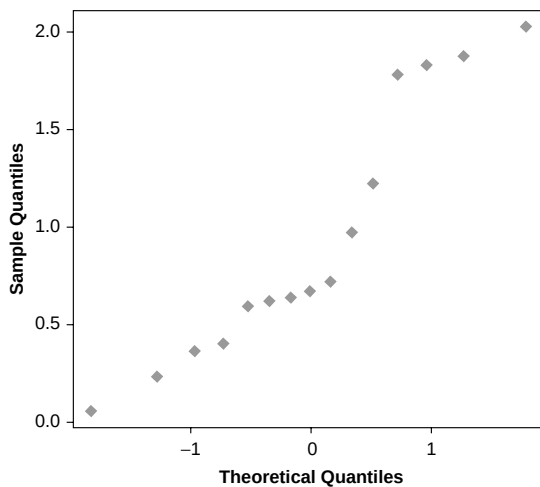


Figure 2 Normal Q-Q Plot

Discussion

Unfortunately, assessing normality through examination of the Q-Q plot can be difficult. It is difficult to determine how much deviation from a straight line is acceptable. Also, small samples ($n < 20$) can have a good deal of variability from linearity, even when the data are known to be from a normal distribution. However, the Q-Q plot can indicate the presence of outliers and the possible reason for departure from univariate normality. When the points fall roughly on a straight line, with the exception of a few points at either edge of the plot, then these points are possible outliers. Rencher gave typical Q-Q plots that one would encounter if the data were heavy-tailed, light-tailed, or positively skewed. D'Agostino, Belanger, and D'Agostino also provided examples of Q-Q plots showing deviations from normality.

—Christopher J. Mecklin

See also Bar Chart; Line Chart; Pie Chart

Further Reading

- D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician*, 44, 316–321.
- Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

NIST/SEMATECH. (n.d.). Quantile-quantile plot. *NIST/SEMATECH e-Handbook of statistical methods* (§1.3.3.24). Retrieved April 12, 2006, from <http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm>

Rencher, A. C. (1995). *Methods of multivariate analysis*. New York: Wiley.

Weisstein, E. W. (n.d.). *Quantile-quantile plot*. Math World: A Wolfram Web Resource. Retrieved April 12, 2006, from <http://mathworld.wolfram.com/Quantile-QuantilePlot.html>

QUALITY OF WELL-BEING SCALE

The Quality of Well-Being Scale (QWB) is a generic measure of health-related quality of life (HRQOL) that combines preference-weighted values for symptoms and functioning. The QWB was developed in the early 1970s and is based on the general health policy model. In the general health policy model, a summary preference-based quality-of-life score is integrated with the number of people affected and the duration of time affected to produce the output measure of quality-adjusted life years (QALYs). QALYs combine morbidity and mortality outcomes into a single number.

In the original QWB, respondents report whether or not each of 27 symptoms were experienced on each of the 6 days prior to the assessment. Functioning is assessed by questions about the presence of functional limitations over the previous 6 days, within three separate domains (mobility, physical activity, and social activity). Unlike measures that ask about general time frames such as “the past 4 weeks” or “the previous month,” the QWB asks whether specific symptoms or functional limitations did or did not occur on a given day. Each symptom complex and functional limitation is weighted using preferences obtained from the ratings of 856 people randomly sampled from the general population. The four domain scores (three functioning, one symptom) are subtracted from 1.0 to create a total score that provides an expression of well-being that ranges from 0 for death to 1.0 for asymptomatic optimal functioning. References on the validation of the instrument are available from the University of California, San Diego, Health Outcomes Assessment Program (see Web site toward the end of

this entry). The questionnaire must be administered by a trained interviewer because it employs a somewhat complex system of questions, including branching and probing. The original questionnaire takes about 15 minutes to complete on average. The authors believe the length and complexity of the original measure has resulted in its underutilization.

In 1996, a self-administered version of the questionnaire was developed to address some of the limitations of the original version. The Quality of Well-Being Scale–Self-Administered (QWB–SA) improves upon the original version in a number of ways. First, the administration of the questionnaire no longer requires a trained interviewer and can be completed in less than 10 minutes. Second, the assessment of symptoms follows a clinically useful review of systems model, rather than clustering symptoms based on preference weights. Third, a wider variety of symptoms are included in the QWB–SA, making it more comprehensive and improving the assessment of mental health.

Both the QWB and QWB–SA are available free of charge to users from nonprofit organizations. A small fee is charged to for-profit users. Information on copyright agreements and user manuals are available at <http://www.medicine.ucsd.edu/fpm/hoap/>.

—Erik J. Groessl and Robert M. Kaplan

See also Life Values Inventory; Measurement; Validity Theory

Further Reading

- Kaplan R. M., & Anderson, J. P. (1996). The general health policy model: An integrated approach. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials* (2nd ed., pp. 309–322). Philadelphia: Lippincott-Raven.
- Kaplan, R. M., Sieber, W. J., & Ganiats, T. G. (1997). The Quality of Well-Being Scale: Comparison of the interviewer-administered version with a self-administered questionnaire. *Psych Health*, 12, 783–791.

are sometimes referred to as *nonrandomized experiments* or *observational studies*). However, quasi-experiments differ from randomized experiments in that units are not randomly assigned to conditions. Quasi-experiments are often used when it is not possible to randomize ethically or feasibly. Therefore, units may be assigned to conditions using a variety of nonrandomized techniques, such as permitting units to self-select into conditions or assigning them based on need or some other criterion. Unfortunately, quasi-experiments may not yield unbiased estimates as randomized experiments do, because they cannot reliably rule out alternative explanations for the effects. To improve causal inferences in quasi-experiments, however, researchers can use a combination of design features, practical logic, and statistical analysis. Although researchers had been using quasi-experiments designs long before 1963, that was the year Donald T. Campbell and Julian C. Stanley coined the term *quasi-experiment*. The theories, practices, and assumptions about these designs were further developed over the next 40 years by Campbell and his colleagues.

Threats to Validity

In 1963, Campbell and Stanley created a validity typology, including *threats to validity*, to provide a logical and objective way to evaluate the quality of causal inferences made using quasi-experimental designs. The threats are common reasons that explain why researchers may be incorrect about the causal inferences they draw from both randomized and quasi-experiments. Originally, Campbell and Stanley described only two types of validity, internal validity and external validity. Thomas D. Cook and Campbell later added statistical conclusion validity and construct validity.

Of the four types of validity, threats to *internal validity* are the most crucial to the ability to make causal claims from quasi-experiments, since the act of randomization helps to reduce the plausibility of many internal validity threats. Internal validity addresses whether the observed covariation between two variables is a result of the presumed cause influencing the effect. These internal validity threats include the following:

QUASI-EXPERIMENTAL METHOD

Quasi-experiments, like all experiments, manipulate treatments to determine causal effects (quasi-experiments

- *Ambiguous temporal precedence*: the inability to determine which variable occurred first, thereby preventing the researcher to know which variable is the cause and which is the effect.
- *Selection*: systematic differences between unit characteristics in each condition that could affect the outcome.
- *History*: events that occur simultaneously with the treatment that could affect the outcome.
- *Maturation*: natural development over time that could affect the outcome.
- *Regression*: occurs when units selected from their extreme scores have less extreme scores on other measures, giving the impression that an effect had occurred.
- *Attrition*: occurs when those who drop out of the experiment are systematically different in their responses from those who remain.
- *Testing*: Repeatedly exposing units to a test may permit them to learn the test, giving the impression that a treatment effect had occurred.
- *Instrumentation*: Changes in the instrument used to measure responses over time or conditions may give the impression that an effect had occurred.
- *Additive and interactive threats to internal validity*: The impact of a threat can be compounded by, or may depend on the level of, another threat.

The other three types of validity also affect the ability to make causal conclusions between the treatment and outcome but do not necessarily affect quasi-experiments more than any other type of experiment. *Statistical conclusion validity* addresses inferences about the how well the presumed cause and effect covary. These threats, such as low statistical power and violation of statistical assumptions, are essentially concerned with the statistical relationship between the presumed cause and effect. *Construct validity* addresses inferences about higher-order constructs that research operations represent. These threats, such as reactivity to the experimental situation (units respond as they want to be perceived rather than to the intended treatment) and treatment diffusion (the control group learns about and uses the treatment), question whether the researchers are actually measuring or manipulating what they intended. *External validity* addresses inferences about whether the relationship holds over variation in persons, settings, treatment variables, and measurement variables. These threats, such as interactions of

the causal treatment with units or setting, determine how well the results of the study can be generalized from the sample to other samples or populations.

Basic Types of Quasi-Experiments

While there are many variations of quasi-experimental designs, basic designs include, but are not limited to, (a) *one-group posttest-only designs*, in which only one group is given a treatment and observed for effects using one posttest observation; (b) *nonequivalent control group designs*, in which the outcomes of two or more treatment or comparison conditions are studied, but the experimenter does not control assignment to conditions; (c) *regression discontinuity designs*, in which the experimenter uses a cutoff score from a continuous variable to determine assignment to treatment and comparison conditions, and an effect is observed if the regression line of the assignment variable on outcome for the treatment group is discontinuous from that of the comparison group at the point of the cutoff; and (d) *interrupted time series designs*, in which many (100 or more) consecutive observations on an outcome over time are available, and treatment is introduced in the midst of those observations to determine its impact on the outcome as evidenced by a disruption in the time series after treatment; and (e) *single-group or single-case designs*, in which one group or unit is repeatedly observed over time (more than twice, but fewer than in a time series), while the scheduling and dose of treatment are manipulated to demonstrate that treatment affects outcome.

The causal logic of threats to validity can also be applied to two other classes of designs that are not quasi-experiments because the cause is not manipulated, as it is in the previous five designs. These are *case-control designs*, in which a group with an outcome of interest is compared with a group without that outcome to see how they differ retrospectively in exposure to possible causes, and *correlational designs*, in which observations on possible treatments and outcomes are observed simultaneously to see if they are related. These designs often cannot ensure that the cause precedes the effect, making it more difficult to make causal inferences than in quasi-experiments.

Design Features

To prevent a threat from occurring or to diagnose its presence and impact on study results, researchers can manipulate certain features within a design, thereby improving the validity of casual inferences made using quasi-experiments. These design features include (a) adding observations over time before (pretests) or after (posttests) treatment to examine trends over time; (b) adding more than one treatment or comparison group to serve as a source of inference about the counterfactual (what would have occurred to the treatment group if they had not received the treatment); (c) varying the type of treatment, such as removing or varying a treatment; and (d) using nonrandomized assignment methods that the researcher can control or adjust, such as using a regression discontinuity design or matching. All quasi-experiments are combinations of these design features, chosen to diagnose or minimize the plausibility of threats to validity in a particular context.

New designs are added to the basic repertoire of designs using these elements. For example, by adding pretest observations to a posttest-only, nonequivalent control group design, existing pretest differences between the treatment and control groups can better be measured and accounted for, which helps reduce effects of selection. Likewise, adding a comparison group to a time series analysis can assess threats such as history. If the outcome for the comparison group varies over time in the same pattern as the treatment outcome, history is a likely threat.

Examples

In 1983 and 1984, the Arizona State Lottery implemented a campaign aimed at increasing the sale of state lottery tickets by offering free tickets to retail customers, who were not asked if they wanted to buy tickets by store clerks. While the researchers examined the effects of the campaign by using various designs, one of the methods used a nonequivalent control group design with pretests and posttests. The 44 retail stores that implemented the campaign were matched with 22 control stores on market shares (the proportion of total state ticket sales made by each retail store for a single game). All stores were measured on their market shares before the campaign

intervention (pretest) and after the program intervention (posttest). Results indicated that there were no differences between the treatment and control groups in market shares at pretest. However, stores that participated in the campaign profited significantly more in market shares than did the control group at posttest.

In 1974, Cincinnati Bell began charging 20¢ per call to local directory assistance and found an immediate and large drop in local directory-assisted calls once this charge was imposed. One hundred eighty monthly observations were collected from 1962 through 1976, which assessed the number of local and long-distance directory-assisted calls. Results of this study found that the number of local and long-distance directory-assisted local calls steadily increased from 1962 (approximately 35,000 calls per day for local calls and 10,000 calls per day for long-distance calls) until 1973 (approximately 80,000 calls per day for local calls and 40,000 calls per day for long-distance calls). However, once the charge for the local calls was imposed, the number of directory-assisted local calls decreased to approximately 20,000 calls per day in 1974. However, the directory-assisted long-distance calls, which did not have a fee imposed, continued to slowly increase over time. This study was an interrupted time series quasi-experiment using a nonequivalent control group (directory-assisted long-distance calls).

Statistical Adjustments

While Campbell emphasized the importance of good design in quasi-experiments, many other researchers sought to resolve problems in making causal inferences from quasi-experiments through statistical adjustments. One such method uses *propensity scores*, the conditional probability that a unit will be in a treatment condition given a set of observed covariates. These scores can then be used to balance treatment and control units on predictor variables through matching, stratifying, covariate adjustment, or weighting. Another method, *selection bias modeling*, attempts to remove hidden bias that occurs when unobserved covariates influence treatment effects by modeling the selection process. A third method uses *structural equation modeling* (SEM) to study causal relationships in quasi-experiments by

modeling latent variables to reduce bias caused by unreliable measures. While these statistical adjustments have been shown to reduce some of the bias present in quasi-experiments, each of these methods has limitations that prevent it from accounting for all of the sources of biased estimates. Therefore, it is often more effective to obtain less biased estimates through good designs than through elaborate statistics.

Conclusion

Quasi-experiments may never rule out threats to internal validity as well as randomized experiments; however, improving the designs can reduce or control for those threats, making causal conclusions more valid for quasi-experiments than they would otherwise be. This can most easily be done by using designs that are most appropriate for the research question and by adding design features to address particular plausible threats to validity that may exist. While certain conditions within field studies may hinder the feasibility of using more sophisticated quasi-experimental designs, it is important to recognize the limitations of designs that are used. In some cases, statistical adjustments can be used to improve treatment estimates; however, even then, causal inferences should be made with caution.

—*M. H. Clark and William R. Shadish*

See also Dependent Variable; Independent Variable; Inferential Statistics

Further Reading

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- McSweeney, A. J. (1978). Effects of response cost on the behavior of a million persons: Charging for directory

assistance in Cincinnati. *Journal of Applied Behavior Analysis*, 11(1), 47–51.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11(6), 691–714.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Trochim, W. M. K. (2002). Quasi-experimental design. In *Research methods knowledge base*. Retrieved June 14, 2005, from <http://www.socialresearchmethods.net/kb/>

QUESTIONNAIRES

The term *questionnaire* denotes a collection of items designed to measure one or more underlying constructs. Given that questionnaires are one of the most widely used research tools in the social sciences, it is not surprising that a large body of literature has developed around various design features in their use.

Method of Administration

Researchers who use questionnaires must first decide what method of administration to employ. One approach is to use a self-administered questionnaire, such as the traditional paper-and-pencil-based booklet completed in a supervised setting or mailed to respondents and completed at their homes or workplaces. Recently, it has become popular to post self-administered questionnaires on Web sites that can be accessed via the Internet. Alternatively, questionnaires can be administered using interviewers to conduct telephone or face-to-face interviews. In choosing one of these methods, researchers should consider the attributes of the project, the possibility of social desirability effects, and the ease of administration.

Self-Administered Questionnaires

Due to their cost-effectiveness and ease of administration, self-administered questionnaires (in either

the traditional or Internet-based form) are popular among social scientists. Because of the sense of anonymity provided by self-administered measures, this method lessens the likelihood of social desirability effects. Thus, these types of questionnaires are especially useful when studying highly sensitive topics. Furthermore, self-administered measures are self-paced, which ensures that respondents have sufficient time to generate meaningful answers to the questions. Internet-based questionnaires tend to be especially cost-efficient, because expenses often associated with a research project (i.e., photocopying of materials and payment of research assistants) are minimized.

Despite such advantages, there are limitations associated with self-administered measures. If researchers mail their questionnaires, they may obtain very low response rates. Furthermore, individuals who take the time to complete the measure may not be representative of the intended sample. Self-administered questionnaires also may not be suitable for respondents with limited cognitive skills or topics that are complex and require extensive explanation.

Interviewer-Administered Questionnaires

One alternative to using self-administered questionnaires is to conduct telephone interviews. Telephone interviews are associated with substantially higher response rates than the use of self-administered questionnaires, which lessens the possibility that one's data will be compromised by nonresponse error. They also allow researchers to probe respondents' answers if they are initially unclear. Unfortunately, there are several drawbacks associated with this method of administration. These include substantially greater expense relative to self-administered questionnaires and increased vulnerability to social desirability effects.

Researchers may also administer their questionnaires via face-to-face interviews. Face-to-face interviews are ideal when one's sample consists of individuals with limited cognitive or verbal abilities, as this type of interview allows researchers to clarify

the meaning of more challenging items through the use of visual props (e.g., show cards). Like telephone interviews, face-to-face interviews also allow researchers to clarify the meaning of ambiguous questions and probe respondents for clarifications to their answers. However, face-to-face interviews are more costly and time-consuming than other methods of administration. Such interviews are also the most vulnerable to social desirability effects, potentially making them an inappropriate method of administration when highly sensitive topics are being studied.

Question Construction

Open- Versus Closed-Ended Questions

One decision that researchers must make in designing the actual questions that make up a survey is whether to use closed-ended questions (which include response alternatives) or open-ended questions (which allow respondents to generate their own answers). Measures containing closed-ended questions may be easier to interpret and complete. However, such questions may fail to provide response options that accurately reflect the full range of respondents' views. While open-ended questions eliminate this problem, they pose other difficulties. To interpret and analyze responses to open-ended questions, researchers must engage in the costly and time-consuming process of developing a coding scheme (i.e., a way to categorize the answers) and training research assistants to use it.

Question Wording

There are several guidelines with respect to question wording that should be followed in generating effective items. Questions should be short and unambiguous, and certain types of items should be avoided. These include "double-barreled questions," which constrain people to respond similarly to two separate issues. Consider the dilemma faced by the respondent who encounters the following question: "How positively do you feel about introducing new social programs, such as government-subsidized day care, to this state?" The respondent might generally harbor

negative feelings about social programs but might view subsidized day care programs quite positively. In light of such conflicting views, this question would be difficult to answer.

It is also worth noting that questions that are substantively identical may yield different answers from respondents, depending on the specific wording used. Consider the following two questions:

Is it appropriate for libraries to *not be allowed* to carry certain books?

Is it appropriate for libraries to be *forbidden* to carry certain books?

Technically, the meaning inherent in these two questions is the same. However, people might respond to them very differently, because “forbidden” has stronger negative connotations than “not allowed.” Thus, subtle variations in wording can have a dramatic effect on responses.

Rating Versus Rank Ordering Objects

If closed-ended questions are to be used, researchers must decide on an appropriate response format. When researchers wish respondents to indicate their relative preferences for a series of objects, they may choose to ask respondents to rank order them, instead of rating each one. The main advantage associated with this approach is that it eliminates the problem of nondifferentiation between objects. Specifically, when evaluating numerous objects using a rating scale, respondents will inevitably be forced to assign the same rating to several of them. If respondents rank order the objects, nondifferentiation ceases to be a problem.

However, there are difficulties associated with this approach. Rank ordering generally does not allow for the possibility that respondents may feel the same way toward multiple objects, and it may compel them to report distinctions that do not exist. Furthermore, respondents may find the process of ranking a large number of objects burdensome. Rank-ordered data can also be difficult to analyze, as it does not lend itself to many statistical techniques commonly used by social scientists.

Order Effects

When researchers employ questions that entail choosing between alternatives, the order of the response options may influence answers. The nature of such order effects depends on the method of administration. If self-administered measures are being used, *primacy effects* (i.e., biases toward selecting one of the first options presented) may pose a threat to the validity of the data. However, if the questionnaire is being orally administered by an interviewer, *recency effects* (biases toward selecting one of the latter options) may occur. Researchers can safeguard against these issues by counterbalancing response options across respondents and testing for order effects.

Number of Scale Points

If researchers employ scale-based measures, they must decide on the number of scale points to include. If an insufficient number of points are used, the measure may be insensitive to variability in respondents' answers. On the other hand, too many scale points may make the differences between points difficult for respondents to interpret, thereby increasing random error. Generally speaking, the optimal length for unipolar scales (measures used to assess the extremity or amount of a construct) is 5 points, while the optimal length for bipolar scales (measures in which the end points reflect opposing responses) ranges from 5 to 7 points.

Labeling Scale Points

Researchers must also decide whether to label the points on their scales. Labeling scale points generally helps respondents to interpret them as intended. However, if the questionnaire is to be administered via telephone, respondents may have difficulty remembering verbal labels. In such instances, numeric labels are preferable.

Including a Midpoint

Another decision is whether to include a midpoint in one's scale. Although midpoints are useful when

there is a meaningful neutral position associated with a question, there are potential disadvantages to including one. Because the midpoint of a scale is often interpreted as reflecting neutrality, people who are not motivated to consider the items carefully may automatically gravitate toward the middle of the scale. Furthermore, the meaning of the midpoint may be somewhat ambiguous. Unless a researcher stipulates what the midpoint signifies, respondents can interpret it in several ways. A midpoint response to an item could potentially indicate ambivalence, neutrality, or indifference on the part of the individual completing the questionnaire. As a result, researchers who fail to label their scales clearly may find it impossible to ascertain what midpoint responses signify.

Nonresponse Options

When people are queried about their attitudes, they may sometimes generate random, spur-of-the-moment responses. This is especially likely when respondents are asked about issues that they think about infrequently. One method of avoiding this problem is to incorporate nonresponse options into one's questions. This allows respondents to indicate that they are unsure of their opinions and alleviates the pressure to generate a substantive response instantaneously. The difficulty associated with this technique is that respondents who actually have opinions about an issue may simply select the nonresponse option if they are not motivated to consider the questions carefully. An alternative approach is to ask respondents to indicate how strongly they feel about their answers to each question (e.g., the certainty of their responses, how important they consider the issue to be). This method requires people to respond to each item, while allowing the researcher to gauge the strength of their answers.

Questionnaire Construction

Thematic Organization Versus Randomization

The way in which researchers structure their questionnaires can have a profound impact on the types of responses that people provide. Researchers can either

order questions by thematic content or randomize them. Often, social scientists have assumed that randomizing their items is more appropriate. However, studies have indicated that organizing items by thematic content makes it easier for respondents to process the content of the questions, thereby reducing random error.

Order Effects

Researchers should also minimize the potential for order effects in their questionnaires. When they complete questionnaire-based measures, people tend to adhere to conversational norms. More specifically, they adhere to the same set of implicit rules that keep regular social interactions running smoothly. For example, respondents avoid providing redundant information when they complete questionnaires. Thus, using questions that overlap in content (e.g., a question pertaining to how satisfied people are with their social lives, followed by a question pertaining to how satisfied people are with their lives in general) may prompt respondents to generate answers that are very different from the ones that they would have given if each question had been posed separately. On a related note, the initial questions in a measure may prime certain concepts or ideas, thereby influencing respondents' answers to subsequent items.

Several steps can be taken to circumvent the problems associated with question order effects. For instance, order effects can be minimized by counterbalancing items. Filler items (i.e., questions that do not pertain to the phenomena being studied) can also be used in order to "isolate" questions that could influence people's responses to subsequent items. Finally, sensitive, controversial questions should be situated at the end of a measure, so that they do not affect respondents' willingness to complete the other items.

In sum, if researchers exercise forethought in designing their questionnaires, the end product of their efforts will likely be a valid measure that is well suited to its purpose.

—*Leandre R. Fabrigar and Anna Ebel-Lam*

See also Measurement; Measurement Error; Reliability Theory; Validity Theory

Further Reading

- Creative Research Systems. (2003). *The survey system*. Retrieved August 23, 2005, from <http://www.surveysystem.com/sdesign.htm>
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Beimer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and quality* (pp. 141–164). Thousand Oaks, CA: Sage.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, CA: Academic Press.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers*. San Francisco: Jossey-Bass.
- Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 223–252). Cambridge, UK: Cambridge University Press.

QUOTA SAMPLING

Quota sampling is a type of survey sampling in which interviewers are directed to gather information from a specified number of members of the population belonging to certain subgroups. The subgroups are sometimes called *strata* or *cells*. In contrast to stratified random sampling, in which interviewers are given specific instructions by the survey planners concerning which individuals to interview, interviewers in quota sampling are given some latitude in selecting the population members to interview. Quota sampling, therefore, is similar to convenience sampling in that the survey designers and planners do not strictly control the selection of the sample. Some control, however, is maintained in the distribution of some sample characteristics.

Suppose an insurance company wants to gather information on a sample of its policyholders. The company's database of clients contains records on the sex, age, and number of policies of all its policyholders. Interviews are to be conducted by telephone. The interviewers are directed to interview 75 individuals in each of 12 strata. The strata are defined by three factors: sex, age, and number of policies. Specifically, the strata are formed by grouping policyholders by sex (female or male), age group (18–35, 36–65, and over 65 years

old), and number of policies (one policy or more than one policy). These instructions will produce a quota sample because the interviewers may call as many customers as needed in order to quickly find people to complete the required interviews. In a stratified random sample, in contrast, the study planners would randomly select 75 individuals and instruct interviewers to gather data from the selected individuals.

Quota samples can be implemented in contexts other than telephone or in-person interviewing. Suppose a researcher wants to know the amount of herbicide per acre applied to agricultural land planted with corn and soybeans in the state of Iowa. The researcher picks 20 counties from around the state. He or she directs data gatherers to collect information on herbicide application for five farms growing corn and five others growing soybeans in the selected counties. If the data gatherers are allowed to choose the most convenient farms or the first farms that will participate, then it is a quota sample. The cells are defined by county and crop. If the researcher randomly selects the farms that the interviewers should visit, then it is not a quota sample.

Quota sampling is an example of *nonprobability sampling*. *Convenience sampling* is another nonprobability sampling scheme. The goal of a survey is to gather data in order to describe the characteristics of a population. A population consists of units, or elements, such as individuals, plots of land, hospitals, or invoices of purchases or sales for a company. A survey collects information on a sample, or subset, of the population. In nonprobability sampling designs, it is not possible to compute the probabilities of selection for the samples overall and, usually, for individuals. The probabilities are unknown, because typically data gatherers are allowed some freedom in selecting convenient units for data collection.

Estimates of population characteristics based on nonprobability samples can be affected by *selection bias*. Since the interviewers choose respondents that they want to interview, there is a potential for selection bias. If the respondents in the survey are systematically different on the variables being measured from the general population, then estimates of characteristics will be different on average from what they would have been with a controlled probability-sampling

scheme. In *probability sampling*, the survey planner or researcher controls which units are in the sample and selects the sample using known probabilities of selection. The probabilities of selection can be used to produce estimates of population characteristics without the problem of selection bias. Examples of probability sampling include simple random sampling, stratified random sampling, and cluster sampling.

Probability sampling is the standard methodology for large-scale surveys intended to support scientific studies and decision making for government policy. Theory supporting the use of probability sampling was developed beginning in the 1930s and has come to be appreciated widely. Quota sampling, on the other hand, is quite common in marketing surveys and less formal studies. Nonprobability sampling certainly can produce useful information for some purposes. One attempt to adjust for the fact that probabilities of selection are unknown is to use *weights*, usually called *survey weights*, in analysis. These weights are computed so that the sum of the weights for sampled individuals in a particular stratum sum to a number proportional to the actual number of people known to exist in the population in the stratum or cell. The cells for which quotas are set are often chosen due to the fact that the proportions of the population in the cells are known. Although this adjustment can help make the sample more representative of the population in analysis, it cannot overcome the fact that there could be a remaining bias due to noncontrolled random selection of the sample.

As an example of weighting or weight adjustment, suppose a bank conducts a quota sample of its account holders by selecting 100 individuals from each of four categories. The four categories are females holding single accounts, males holding single accounts, females holding multiple accounts, and males holding multiple accounts. If the hypothetical proportions of all account holders in these four categories are as given in Table 1, then the sample that contains 25% of

Table 1 Quota Sample of Account Holders

Percentage in the Population	Number of Accounts		
	One Account	More Than One	Total
Sex of Account Holder			
Female	21%	27%	48%
Male	24%	28%	52%
Total	45%	55%	100%

Table 2 Weights

Weights	Number of Accounts	
	One account	More than one
Sex of account holder		
Female	$25/21 = 1.19$	$25/27 = 0.93$
Male	$25/24 = 1.04$	$25/28 = 0.89$

the respondents in each category is slightly out of balance relative to the population counts.

Weights can be computed as the fraction of the sample in a category divided by the fraction of the population in the same category. The weights in this example are given in Table 2. The bank is interested in the amount of credit card debt held by its account holders. Since the credit cards might not be issued through the bank, the bank cannot simply compute the amount of debt; it must ask members of the sample for the information. A direct average of the amount of debt among the members of the sample might not represent the population very well, because the proportions by category are not exactly equal to those in the population. Instead, a weighted average using weights given in Table 2 might be better.

Of course, the selection bias produced by the quota sample might still lead to a systematic overestimate or underestimate of credit card debt.

The method of *focus groups* is another method of collecting information that can be seen as being similar in spirit to quota sampling. In focus groups, researchers meet with representatives of a population, ask questions, and hold discussions on topics of interest. Researchers conducting focus group studies and interviews with key informants about a population hope to gain insight into issues, concerns, and terminology relevant to the population's perception of an issue. Quota

sampling, instead of bringing a group together for an in-depth discussion, gathers data using a survey from a sample of individuals. A substantial research project might use several methods of gathering data, including focus groups, a pilot survey or a preliminary quota survey, and then a large-scale probability sample survey.

—Michael D. Larsen

See also Convenience Sampling; Nonprobability Sampling; Probability Sampling

Further Reading

- Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.
- Curtice, J., & Sparrow, N. (1997). How accurate are traditional quota opinion polls? *Journal of the Market Research Society*, 39(3), 433–438.

- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.
- Lynn, P., & Jowell, R. (1996). How might opinion polls be improved? The case for probability sampling. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 159, 21–28.
- Owen, L., McNeill, A., & Callum, C. (1998). Trends in smoking during pregnancy in England, 1992–7: Quota sampling surveys. *British Medical Journal*, 317, 728–730.
- Statistics Canada. (n.d.). *Statistics: Power from data!* Retrieved April 12, 2006, from http://www.statcan.ca/english/edu/power/ch13/non_probability/non_probability.htm

R

Scientists should always state the opinions upon which their facts are based.

—Author Unknown

RANDOM NUMBERS

Random numbers are useful for a variety of purposes, such as generating data encryption keys, simulating and modeling complex phenomena, and selecting random samples from larger data sets. They have also been used aesthetically (e.g., in literature and music) and are, of course, ever popular for games and gambling. When discussing single numbers, a random number is one that is drawn from a set of possible values, each of which is equally probable (i.e., a uniform distribution). When discussing a sequence of random numbers, each number drawn must be statistically independent of the others.

There are two main approaches to generating random numbers using a computer. Pseudorandom number generators (PRNGs) are algorithmic approaches that use mathematical formulae (e.g., the linear congruential method) or simply precalculated tables to produce sequences of numbers that appear random. PRNGs are efficient and deterministic, meaning that they can produce many numbers in a short time and a given sequence of numbers can be reproduced at a later date if the starting point in the sequence is known. PRNGs are typically periodic, meaning that the sequence will eventually repeat itself. These

characteristics make PRNGs suitable for applications where many numbers are required and where it is useful that the same sequence can be replayed easily, such as in simulation and modeling applications. PRNGs are not suitable for applications where it is important that the numbers be really unpredictable, such as data encryption and gambling.

In comparison, true random number generators (TRNGs) extract randomness from physical phenomena, such as quantum events or chaotic systems. For example, random numbers can be generated by measuring the variations in the time between occurrences of radioactive decay (quantum events) or the variations in amplitude of atmospheric noise (caused by the planet's chaotic weather system). TRNGs are generally much less efficient than PRNGs, taking a considerably longer time to produce numbers. They are also nondeterministic, meaning that a given sequence of numbers cannot be reproduced, although the same sequence may, of course, occur several times by chance. TRNGs have no period. These characteristics make TRNGs suitable for roughly the set of applications for which PRNGs are unsuitable, such as data encryption, games, and gambling. Conversely, the poor efficiency and nondeterministic nature of TRNGs make them less suitable for simulation and modeling applications.

Many statistical tests exist that can be used to assess the randomness of numbers generated with either approach. Examples include the Chi-Square Test, the Run Test, the Collision Test, and the Serial Test. However, testing for randomness is not straightforward, because each possible sequence is equally likely to appear, and good random number generators therefore will also produce sequences that look non-random and fail the statistical tests. Consequently, it is impossible to prove definitively whether a given sequence of numbers (and the generator that produced it) is random. Rather, random numbers from a given generator are subjected to an increasing number of tests, and as the numbers pass more tests, the confidence in their randomness (and the generator that produced them) increases.

—Mads Haahr

See also Chance; Monte Carlo Methods; Random Sampling

Further Reading

Knuth, D. E. (1997). *The art of computer programming, Volume 2: Seminumerical algorithms* (3rd ed.). Reading, MA: Addison-Wesley.

The National Institute for Standards and Technology's guidelines for true and pseudorandom number generation and testing: <http://csrc.nist.gov/rng/>

Theory and practice of pseudorandom number generation: <http://random.mat.sbg.ac.at/>

True random number service based on atmospheric noise: <http://www.random.org/>

True random number service based on radioactive decay: <http://www.fourmilab.ch/hotbits/>

RANDOM SAMPLING

Three concepts that are relevant to understanding “random sampling” are population, sample, and sampling. A *population* is a set of elements (e.g., scores of a group of prisoners on a personality scale, weights of newborns of mothers younger than 18, longevity of smokers in New York City, and changes in the systolic blood pressure of hypertensive patients exposed to relaxation training) defined by the researcher's

interests. Populations can be defined narrowly or broadly and can consist of a few or many elements. For example, a researcher concerned with effects of relaxation training on systolic blood pressure might be interested only in pre- to posttreatment systolic blood pressure changes (i.e., pre–post) in the small group of patients included in her study, or she might be interested in the efficacy of relaxation training in the collection of all persons who could be classified as “elderly obese hypertensive patients” by some operational definitions of these terms.

A *sample* is a subset of a population, and *sampling* refers to the process of drawing samples from a population. Sampling is generally motivated by the unavailability of the entire population of elements (or “data”) to the researcher and by her interest in drawing inferences about one or more characteristics of this population (e.g., the mean, μ , or variance, σ^2 , of the population). Thus, in the hypertension example, the researcher might be interested in estimating from her sample data the mean (μ) or variance (σ^2) of pre- to posttreatment changes in systolic blood pressure of the population consisting of persons who fit the operational definition of “elderly obese hypertensive patients.” She might also be interested in using the sample data to test (the tenability of) some hypothesis about μ , such as the hypothesis that the average change in systolic blood pressure (μ) in the population of hypertensives exposed to relaxation training is zero. This hypothesis will be called the “null hypothesis” and will be abbreviated “ H_0 ” below. Because characteristics of a population, such as μ and σ^2 , are defined as “parameters” of the population, the principal objectives of random sampling can often be described in terms of (a) estimation of population parameters and (b) testing of hypotheses about population parameters. Parameters that are often of interest are means (μ s) and linear combinations of means that provide information about mean differences, trends, interaction effects, and so on.

Simple random sampling refers to a method of drawing a sample of some fixed size n from the population of interest, which ensures that all possible samples of this size (n) are equiprobable (i.e., equally likely to be drawn). Sampling can be carried out with

replacement or without replacement. Sampling is with replacement if, in the process of drawing the n elements of the sample consecutively from the population, each element is replaced in the population prior to drawing the next element. For example, if the population of interest consisted of Minnesota Multiphasic Personality Inventory Depression scores of 60, 65, and 70, drawing a sample of size $n = 2$ with replacement involves replacement of the first-drawn score prior to making the second draw. Sampling is without replacement in this example if the first-drawn score is not replaced prior to making the second draw. Therefore, there would be nine possible samples of size $n = 2$ (in this example) when sampling was with replacement, as shown below:

60,60	60,65	60,70
65,60	65,65	65,70
70,60	70,65	70,70

and six possible samples of size $n = 2$ if sampling is without replacement, as shown below:

	60,65	60,70
65,60		65,70
70,60	70,65	

If, in this example, the sampling is not only with replacement but also random, the probability or likelihood of drawing any one of the nine possible samples [e.g., (60,60), (60,65)] would be equal to that of drawing any of the other eight possible samples (i.e., $1/9 = .111$). However, if the sampling is random and without replacement, the probability of drawing any one of the six samples that are possible without replacement [e.g., (60,65), (60,70)] would be equal to that of drawing any of the other five possible samples (i.e., $1/6 = 0.167$).

Random Sampling and Representativeness

Authors sometime state or imply that random sampling generates samples that are representative of the populations from which they have been drawn. If this is true, then two random samples of size n drawn from the same population should be very similar. But

in fact, randomly drawn samples can have a non-negligible probability of being very unrepresentative of the population from which they were drawn, and two random samples drawn from the same population can be very dissimilar, especially when these samples are small. For example, suppose that relaxation training was ineffective in the sense that the average (pre-post) systolic blood pressure change was zero ($\mu = 0$) in the population of interest, and assume that the distribution of blood pressure changes of this population was actually bell-shaped (normal, Gaussian) with a standard deviation of $\sigma = 22$ mm Hg. Then it can be shown (using the central limit theorem) that (a) the mean (M) of a *random sample* of $n = 4$ drawn from this population would have the nonnegligible probability of .36 of under- or overestimating the population mean (i.e., $\mu = 0$) by more than 10 mm, and (b) a *random sample* of that size would have a probability of 0.18 of having a mean $M > 10$ mm. Furthermore, two independently drawn random samples of this size could, with nonnegligible probability, yield very different sample means.

Suppose that the probability that the two sample means differ by more than 10 mm Hg is 0.52. Because systolic blood pressure changes of 10 or more mm are associated with important health benefits or risks (for example, it has been estimated that every 10 mm Hg increase in systolic blood pressure is associated with a 30% increase in the risk of heart attack), nonnegligible overestimation or underestimation of effects of relaxation training on systolic blood pressure, as well as important nonequivalence of samples with respect to average blood pressure changes, could occur quite frequently with random samples of that size. Clearly, random sampling need not ensure representativeness or equivalence of samples. However, random sampling generally permits drawing useful inferences about population characteristics from sample data, regardless of the sizes of the samples (as illustrated below).

Random Sampling and the Estimation of Parameters

When a sample of size n is drawn at random from a population with mean μ , the mean of that sample (M)

will be an unbiased estimator of μ . This means that if this sampling “experiment” (the drawing of the random sample) was repeated an infinitely large number of times, the mean (also called the “expected value”) of the M s would be equal to μ . Thus, if the mean of change scores in a population of hypertensive patients who were exposed to relaxation training was in fact $\mu = 5$ mm, and if independent random samples of size 4 were to be drawn an infinitely large number of times from this population, the mean or expected value of the sample means would be 5 mm; in this situation, random sampling would result in “absence of bias” of the statistic M in the sense that the average M would be equal to 5. The sample mean M would be described as a “point estimate” of the parameter μ . [The expression “point estimate” refers to the fact that the M of a sample is a single value or point, as opposed to a range of values or intervals (see below).]

Given that the distribution from which the sample of size n has been randomly selected is normal and that its mean is μ and its standard deviation is σ , it can be shown that an interval that is defined as extending from $[M - z_{.025} \sigma/n^{.5}]$ to $[M + z_{.025} \sigma/n^{.5}]$ has a probability of 0.95 of including μ (where $z_{.025}$ is 1.96, the 97.5th centile of the standard normal distribution). This interval estimate of μ is called a two-sided 95% confidence interval. The probability that a 95% confidence interval constructed in this way includes the parameter that it is used to estimate is 0.95. For example, given that $n = 49$, $\sigma = 22$, and $M = 10$, the 95% confidence interval would extend from $[10 - 1.96(22/7)]$ to $[10 + 1.96(22/7)]$, that is, from 3.84 to 16.16. The researcher could be quite sure or confident that the average drop in systolic blood pressure in the population of hypertensives would be somewhere between 3.84 and 16.16. However, if the sample size had been $n = 4$, the same sample mean of $M = 10$ would yield a much larger 95% confidence interval. With $n = 4$, $\sigma = 22$, and $M = 10$, the 95% confidence interval would extend from $[10 - 1.96(22/2)]$ to $[10 + 1.96(22/2)]$, that is, from -11.56 to 31.56. A much wider range of values of μ would be consistent with the sample data. It should be noted that in these examples, the

construction of the confidence intervals required knowledge not only of sample data but also of σ , a parameter that is often unavailable to the researcher. However, given random sampling, confidence intervals for μ s can also be constructed (using methods discussed in most statistic textbooks) without knowledge of σ . These examples illustrate the fact that, irrespective of the representativeness of random samples, random sampling generally permits interpretable inferences about parameters of interest to the researcher.

Random Sampling and Parametric Hypothesis Testing

Testing a hypothesis about the value of a parameter (e.g., $H_0: \mu = 0$) can be viewed as determining whether this value is or is not consistent with the sample data. Thus, a researcher who tests $H_0: \mu = 0$ in the above example in effect determines whether $\mu = 0$ is or is not plausible given the observed sample data. What is apparent from this perspective on hypothesis testing is that confidence intervals, whose construction assumes random sampling, can provide the information needed for hypothesis testing by explicitly specifying a range of values of the parameter that would be plausible given the sample data. Thus, the fact that the 95% confidence interval for the sample of $n = 49$ excluded zero would usually lead to a decision to reject the null hypothesis that $\mu = 0$ because this interval did not include zero. On the other hand, the fact that the 95% confidence interval for the sample of $n = 4$ included zero would generally lead to the conclusion that $\mu = 0$ was consistent with the sample data—that is, consistent with the possibility that the average change in systolic blood pressure of patients in the population could be zero. Clearly, because the methods used to construct the confidence intervals assume random sampling, the hypothesis tests based on information yielded by the confidence intervals also require tenability of the random sampling assumption. Random sampling implies models that make possible both estimation (point and interval estimation) and hypothesis testing about parameters of interest to researchers.

Random Sampling Versus Randomization

Random sampling, as noted above, refers to a method of drawing a sample of size n from a population, which ensures that all possible samples of that size are equiprobable. Randomization, on the other hand, refers to a method of assigning or allocating $n = n_1 + n_2 + \dots + n_k$ participants in a study to k groups where n_1 subjects are assigned to Group 1, n_2 subjects to Group 2, . . . , and n_k subjects to Group k , which ensures that all of the possible ways of making the assignments are equiprobable. For example, with $n = 5$ participants (ABCDE) and two groups, there would be 10 ways of making the assignments so that $n_1 = 2$ and $n_2 = 3$: (AB,CDE), (AC,BDE), (AD,BCE), (AE,BCD), (BC,ADE), (BD,ACE), (BE,ACD), (CD,ABE), (CE,ABD), (DE,ABC); random assignment or randomization would refer, in this example, to a method that would ensure equiprobability ($1/10 = 0.10$) of the 10 possible allocations. Randomization is feasible with a “sample of convenience,” that is, a sample of n available subjects all of whom will be used in a study, and inferential statistical methods (e.g., hypothesis testing) can be used both with random sampling (from populations whose characteristics are of interest to researchers) and with random assignment of subjects from a sample of convenience (i.e., randomization). However, although random sampling allows statistical inferences to be drawn (from a study’s sample data) about characteristics of populations from which random samples have been drawn, randomization (without random sampling) allows statistical inferences only about the participants of the study. Nevertheless, nonstatistical inferences about individuals not included in the study are often possible with randomization.

—Louis M. Hsu

See also Sample

Further Reading

- Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, 66, 485–487.
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart and Winston.

- Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, 57, 131–137.
- Hunter, M. A., & May, R. B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Canadian Journal of Experimental Psychology*, 57, 176–188.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Erlbaum.
- Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, 4, 117–126.
- Siemer, M., & Joorman, J. (2003). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods*, 8, 497–517.
- Stilson, D. W. (1966). *Probability and statistics in psychological research and theory*. San Francisco: Holden-Day.

Random sampling versus randomization—different implications: <http://www.tufts.edu/~gdallal/rand.htm>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Gershater-Molko, R. M., Lutzker, J. R., & Wesch, D. (2002). Using recidivism to evaluate Project SafeCare: Teaching bonding, safety, and health care skills to parents. *Child Maltreatment*, 7(3), 277–285.

Project SafeCare was an in-home research and intervention, grant-funded program designed to teach parents who were reported for child abuse and neglect. Parents who participated in Project SafeCare received training in three aspects of child care: treating illnesses and maximizing their health care skills (health), positive and effective parent-child interaction skills (bonding), and maintaining hazard-free homes (safety) for their children. True **random sampling** was used to select families for participation. Postcontact (after initial intake was made and the program began) incidents of child abuse and neglect for maltreating parents who participated in and completed Project SafeCare were compared to a comparison group of maltreating families from the point of initial intake through a 24-month follow-up period. The comparison group (the Family Preservation group) received intervention from Family Preservation programs. Families who participated in Project SafeCare had significantly lower reports of child abuse and neglect than did families in the comparison group.

RANGE

Among all the measures of variability, the range is the most general and is an overall picture of how much variability there is in a group of scores. It provides an impression of how far apart scores are from one another and is computed by simply subtracting the lowest score in a distribution from the highest score in the distribution.

In general, the formula for the range is

$$R = H - L,$$

where

R is the range,

H is the highest score in the data set,

L is the lowest score in the data set.

Take the following set of scores, for example (shown here in descending order):

98, 86, 77, 56, 48

In this example, $98 - 48 = 50$. The range is 50.

Although the formula you see above is the most common one for computing the range, there is another definition and formula as well. The first is the exclusive range (as shown above), which is the highest score minus the lowest score (or $h - l$) and the one that was just defined. The second kind of range is the inclusive range, which is the highest score minus the lowest score plus 1 (or $h - l + 1$). The exclusive range is the most common, but the inclusive range is also used on occasion. There are other types of ranges as well, such as the semi-interquartile range.

The range is used almost exclusively to get a very general estimate of how wide or different scores are from one another—that is, the range shows how much spread there is from the lowest to the highest point in a distribution. So, although the range is fine as a general indicator of variability, it should not be used to reach any conclusions regarding how individual scores differ from one another.

Shortcomings of the range as a measure of variability are two. First, because the range reflects the difference between only two numbers (and not any others), the only thing we know about the set of scores is these two numbers and nothing about potential outliers or scores that are far beyond the general characteristics of the distribution. For example, a range of 60 (100–40) may very well include only one number above 80 (which is 100), while the same range might reflect a distribution of scores with 20 data points above 80 (83, 92, and 97, for example). Second, because only two numbers are used, there is no discerning the shape or meaning or patterns of a set of data. Knowing the range tells us nothing about the shape of a distribution.

—Neil J. Salkind

Further Reading

Zimmerman, D. W. (2000). Restriction of range and correlation in outlier-prone distributions. *Applied Psychological Measurement*, 24(3), 267–280.

Range and variability: http://psych.rice.edu/online_stat/chapter3/variability.html

RASCH MEASUREMENT MODEL

A Rasch measurement model is an example of additive, conjoint, fundamental measurement by which one can create linear, objective measures applicable to the human sciences (such as in, but not limited to, education, psychology, medicine and health, marketing and business, and judging in sports). Rasch measurement models show how to determine what is measurable on a linear scale, how to determine what data can be used reliably to create a linear scale, and what data cannot be used in the creation of a linear scale. In a linear scale, equal differences between the numbers on the scale represent equal amounts of the measure.

Raw Scores to Linear Measures

Rasch measurement models can be used to convert many different types of raw score data to a linear

scale, such as dichotomous data; rating response scores; partial credit scores; skills or achievement judging scores (from, say, 1 to 10); essay marks; partial skills development; quality levels or levels of success; and so on. They can be applied to achievement data (as in subjects at schools and universities); attitude data; personality data; quality of life data; levels of health data; behavioral data; data given by groups of judges on tasks or skills (such as in diving, dancing, and ice skating); and so on.

“Scale-Free” Measures and “Sample-Free” Item Difficulties

An important point to understand is that when the data fit a Rasch measurement model, the differences between the person measures and the item difficulties can be calibrated together in such a way that they are freed from the distributional properties of the incidental parameter, because of the mathematics involved in the measurement model. This means that “scale-free” measures and “sample-free” item difficulties can be estimated with the creation of a mathematically objective linear scale with standard units. The standard units are called logits (the log odds of successfully answering the items).

A requirement for measurement is that the units should be the same size across the range of the variable measures, and this is not true with percentage scores, or summed scores from a set of achievement or attitude items, where small changes in the probability of success are related to large changes in person abilities at the bottom and top of percentage scales, all of which are nonlinear. By converting the probability of success to log odds and logits as the unit in Rasch measurement, the nonlinear problem is greatly reduced, particularly at the top of the scale.

The Simple Logistic Model of Rasch

The simplest Rasch measurement model for creating a linear scale was developed by Georg Rasch (1901–1980) and published in 1960. The simple logistic model (SLM) of Rasch has two parameters: one representing a measure for each person on a variable

and the other representing the difficulty for each item (it is sometimes called the *one-parameter model* in the literature).

Requirements of the SLM of Rasch

1. Items are designed to be conceptually ordered by difficulty along an increasing continuum from easy to harder for the variable being measured.

2. In designing the items (using three as an example), one keeps in mind that person measures of the variable are conceptualized as being ordered along the continuum from low to high according to certain conditions. The conditions in this example are that persons with low measures will have a high probability of answering the easy items positively, and a low probability of answering the medium and hard items positively. Persons with medium measures will have a high probability of answering the easy and medium items positively, and a low probability of answering the hard items positively. Persons with high measures will have a high probability of answering the easy, medium, and hard items positively. These conditions are tested through a Rasch analysis.

3. Data are collected from persons on the items and scored dichotomously (0/1 or 1/2), as in, for example, wrong/right, no/yes, none/a lot, or disagree/agree.

4. Each item is represented by a number, estimated from the data that represent its difficulty (called an item parameter in the mathematical representation of the Rasch model) that does not vary for persons with different measures of the variable. Persons with different measures responding to the items have to agree on the difficulty of the items (such as easy, medium, and hard, as used in this example). If the persons do not agree on an item’s difficulty, then this will be indicated by a poor fit to the measurement model, and then the item may be discarded as not belonging to a measure on this continuum.

5. Each person is represented by a number, estimated from the data that represent his or her measure of the variable (called a person parameter in the mathematical representation of the Rasch model) that does

not vary for items of different difficulty along the continuum. If different items do not produce agreement on a person measure, then this will be indicated by a poor fit to the measurement model, and then one examines the person response pattern (and the items).

6. Rasch measurement models use a probability function that allows for some variation in answering items such that, for example, a person with a high attitude measure sometimes might give a low response to an easy item, or a person with a medium achievement measure sometimes might answer a hard item correctly. If the person response pattern shows too much disagreement with what is expected, then it may be that the person has not answered the items properly or consistently, and that person's results may be discarded, or the item may be too hard or too easy.

Equations for the Simple Logistic Model of Rasch

$$\begin{array}{l} \text{Probability of answering} \\ \text{positively (score 1)} \\ \text{for person } n \end{array} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

$$\begin{array}{l} \text{Probability of answering} \\ \text{negatively (score 0)} \\ \text{for person } n \end{array} = \frac{1}{1 + e^{(B_n - D_i)}}$$

where

e is the natural logarithm base ($e = 2.7318$),

B_n is the parameter representing the measure (ability, attitude, performance) for person n ,

D_i is the parameter representing the difficulty for item i .

These equations are solved from the data by taking logarithms and applying a conditional probability routine with a computer program such as the Rasch Unidimensional Measurement Models (RUMM), Winsteps, or ConQuest. The RUMM program is currently the best of the main computer programs for Rasch measures for two reasons. One is that the RUMM program provides a comprehensive set of

output data to test many aspects of the conceptual model of the variable, the answering consistency of the response categories, both item and person fit to the measurement model, and targeting. And second, the RUMM program produces a wonderful set of colored, graphical maps for many aspects of the measurement.

The Partial Credit Model of Rasch

The partial credit model (PCM) of Rasch can be thought of as an extension of the SLM from two response categories to three or more response categories or outcomes. The conditions, requirements, and output of the PCM are similar to the SLM, except that there are now more item parameters and more item output, and the equations are more complicated. The PCM can be applied to any set of data scored, judged, or answered in three or more ordered outcome categories where the level of outcome is conceptualized on a continuum from low to high. Seven examples of data that can be analyzed with the PCM are listed.

1. Ordered levels of knowledge, understanding, or skills, such as school tests in which questions (items) are marked in a set of categories, such as 0, 1, 2, 3, 4, 5, depending on the level of understanding displayed. One might have a standard for each of the six levels to help with consistency of scoring.
2. Ordered levels of tasks in problem solving where, for example, completion of a complex problem requires three intermediate ordered tasks to be completed, so the problem is scored 0, 1, 2, 3, 4, depending on the level completed.
3. Ordered levels of health after an operation where, for example, patients describe their level of well-being on a scale from 1 (*unable to eat and move without pain*) to 10 (*able to eat and walk around without discomfort*)
4. Rated level of marks, for example, by one teacher assigning scores of 0 to 10 on a task such as writing an essay or completing an assignment
5. Respondents' self-ratings on attitude, personality, or behavior items using, for example, a unipolar set of responses such as *no agreement* (0), *weak agreement* (1), *strong agreement* (2), and *very strong agreement* (3), or using a bipolar set of responses such as *strongly disagree* (0), *disagree* (1), *agree* (2), *strongly agree* (3)

6. A single judge, teacher, or coach providing a judgment on the level of skill achieved in a sport or activity
7. A computerized test item that is programmed to provide a series of ordered feedback hints to allow further chances to answer correctly and ordered scores given accordingly

Equations for the Partial Credit Model of Rasch

$$\text{Probability of person } n \text{ scoring in outcome category } x \text{ of item } i = \frac{e^{\sum_{j=1}^x (B_n - \delta_{ij})}}{1 + \sum_{k=1}^{M_i} e^{\sum_{j=1}^k (B_n - \delta_{ij})}} \quad (\text{for } x = 1, 2, 3, 4, \dots, M_i)$$

$$\text{Probability of person } n \text{ scoring in outcome category } x \text{ of item } i = \frac{1}{1 + \sum_{k=1}^{M_i} e^{\sum_{j=1}^k (B_n - \delta_{ij})}} \quad (\text{for } x = 0)$$

where

e is the natural logarithm base ($e = 2.7318$),

$\sum (B_n - \delta_{ij})$ is the sum of $B_n - \delta_{ij}$,

B_n is a parameter representing the measure (ability, attitude, skill or performance) for person n ,

$\delta_{i1}, \delta_{i2}, \delta_{i3}, \dots, \delta_{iM_i}$ are a set of parameters for item i that jointly locate the model probability curves for item i . There are M_i item parameters for an item with $M_i + 1$ outcome categories.

Rasch Data Analysis With the RUMM Computer Program

Using the PCM of Rasch with the RUMM computer program, there are eight data analysis tests (output) provided in the creation of a linear, unidimensional scale. This output is similar for the PCM, the ELM, the RRM, and the SLM (except that for the SLM, there are no ordered thresholds, just one threshold).

1. *Testing that the response categories are answered consistently and logically.* The RUMM program does this with two outputs: (a) It calculates

threshold values between the response categories for each item (where there are odds of 1:1 of answering in adjacent categories), and (b) it provides response category curves showing the graphical relationship between the linear measure and the probability of answering each response category.

2. *Testing for dimensionality.* An item-trait test of fit is calculated as a chi-square with a corresponding probability of fit. It tests the interaction between the responses to the items and the person measures along the variable and shows the collective agreement for all items across persons of different measures along the scale. If there is no significant interaction, one can infer that a single parameter can be used to describe each person's response to the different item difficulties, and thus we have a unidimensional measure.

3. *Testing for good global item-person fit statistics.* The item-person test of fit examines the response patterns for items across persons, and the person-item test of fit examines the response patterns for persons across items using residuals. Residuals are the differences between the actual responses and the expected responses as estimated from the parameters of the measurement model. When these residuals are summed and standardized, they will approximate a distribution with a mean near zero and standard deviation near one, when the data fit a Rasch measurement model.

4. *Person Separation Index.* Using the estimates of the person measures and their standard errors, the RUMM program calculates a Person Separation Index that is constructed from a ratio of the estimated true variance among person measures and the estimated observed variance among person measures. This tests whether the standard errors are much smaller than the differences between the person measures.

5. *Testing for good individual item and person residuals.* Residuals are the differences between the observed values and the expected values estimated from the parameters of the Rasch measurement model. It is instructive to examine these outputs because they give an indication of whether persons are answering items in a consistent way, and they also give an indication of individual person and individual item fit to the measurement model.

6. *Item characteristic curves.* Item characteristic curves examine how well the items differentiate between persons with measures above and below the item location. It also shows a comparison between the observed and expected proportions correct for a number of class intervals of persons.

7. *Person measure/item difficulty map.* The RUMM program produces two types of person measure/item difficulty maps. These maps show how the person measures are distributed along the variable and how the item difficulties are distributed along the same variable (measured in logits). They show which items are easy, which ones are of medium difficulty, and which ones are hard. They show how well the item difficulties are targeted at the person measures. That is, they show whether the items are too easy or too hard for the persons being measured and whether new items need to be added, or whether there are too many items of similar difficulty (some of which are thus not needed).

8. *Testing for construct validity.* Suppose that your items are conceptually ordered by increasing difficulty (downward) and the perspectives are ordered by increasing difficulty (to the right), and this represents the structure behind your variable. In Rasch measurement, all the item difficulties are calculated on the same linear scale, and so the item difficulties can be compared with their conceptualized order. In this case, the item difficulties increase vertically downward for each perspective by item and they increase horizontally to the right for each item by perspective. This provides strong support for the structure of the variable because it was postulated before the data were collected and analyzed.

The Many-Facets Model of Rasch

The Many-Facets Model of Rasch (MFMR) applies to situations where judges rate persons on, for example, skills, items, essays, or behaviors. The model states that persons have an estimated probability of being given any rating on an ordinal scale (e.g., 1–10, as given by judges of diving) on any item by any judge. Each observation is modeled by an ordinal rating from a loglinear combination of the measures of one element from the appropriate facets, for example, a person performance

from the person ability facet, a judge severity from the judge facet, and an item difficulty from the item facet. True to Rasch measurement, the measures are estimated from the data but are statistically independent of observations comprising the data. These measures form a linear scale on which the elements of all the facets are located. Estimates of fit to the measurement model can be made by examining the differences between the actual observations and the predicted observations estimated from the parameters of the measurement model, as is done in the other Rasch models.

An Equation for a Many-Facets Model of Rasch

The following equation applies to an example where persons are rated by judges on three or more items, each of which has its own ordinal rating scale:

$$\text{Ln} (P_{nijk}/P_{nijk} - 1) = B_n - D_i - C_j - F_{ik},$$

where

Ln is the natural logarithm (to the base e),

P_{nijk} is the probability that person n will gain a rating of category k from judge j on item i ,

B_n is a parameter representing the ability, skill, or performance of person n ,

D_i is a parameter representing the difficulty of item i ,

C_j is a parameter representing the severity of judge j ,

F_{ik} is a parameter representing the additional difficulty of being rated in category k beyond that of being rated in category $k - 1$ for item i ,

K_i is the highest rating scale category. Zero is the lowest.

Some Special Applications of Rasch Measurement

The creation of linear measures in the human sciences through Rasch measurement and the development of personal computers has allowed for the use of two special applications: item banks and computerized adaptive testing. Item banks involve the collection of thousands of school achievement items relating to a topic, such as algebra, that have been calibrated on a linear scale. Teachers can choose a set of items on the

topic, calibrated from easy to hard, and use the items to test their students. If an appropriate computer program has been created, students can use a computer to select items appropriate to their level of understanding and attempt only those. Because all the items have been calibrated on the same linear scale, students (and their teachers) can obtain a measure of their level of understanding.

—Russell F. Waugh

Further Reading

- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.
- Andrich, D. A. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, 1(4), 363–378.
- Andrich, D. A., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17(3), 253–276.
- Andrich, D. A., Sheridan, B. E., & Luo, G. (2006). *A Windows-based item analysis program employing Rasch unidimensional measurement models*. Perth, Western Australia: RUMM Laboratory.
- Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. Maple Grove, MN: JAM.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Garner, M., Engelhard, G., Jr., Wilson, M., & Fisher, W., Jr. (Eds.). (2005). *Advances in Rasch measurement*. Maple Grove, MN: JAM Press.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA.
- Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, 1(4), 279–298.
- Rasch, G. (1992). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: MESA. (Original work published 1960)
- Smith, E., & Smith, R. (Eds.). (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.
- Smith, E., & Smith, R. (Eds.). (2005). *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: JAM.
- Waugh, R. F. (Ed.). (2003). *On the forefront of educational psychology*. New York: Nova Science.
- Waugh, R. F. (Ed.). (2005). *Frontiers in educational psychology*. New York: Nova Science.

RATIO LEVEL OF MEASUREMENT

The most precise of the four levels of measurement is called the ratio level. The ratio scale is unique because, unlike the interval scale, it contains an absolute rather than arbitrary zero. In other words, a score of 0 on a ratio-level measure indicates a complete absence of the trait or construct being measured. Whereas at a temperature of 0° on a Fahrenheit scale, some molecules are still moving (because the zero of the scale is arbitrary), a reading of 0° on a Kelvin temperature scale signifies a total lack of molecular movement. Therefore, the Fahrenheit scale is an example of interval measurement, whereas the Kelvin scale is considered ratio-level measurement.

The reason the absolute zero makes the ratio scale unique is that it allows us to make meaningful fractions. On the Kelvin scale, 20° is twice as “hot” (has twice the amount of moving molecules) as 10°, and 150° is three times as “hot” as 50°. Using the Fahrenheit scale, we cannot make such a comparison; 20° is not twice as hot as 10° because the zero point on the scale does not indicate a complete absence of molecular movement and because temperatures below 0° are possible.

Whereas ratio-level data are fairly common in physical sciences (zero molecular movement, zero light, zero gravity), they are rarer in behavioral and social science fields. Even if someone scores a zero for whatever reason on an IQ test, the tester would not declare that the test taker has no intelligence. If a student receives 0 points on a vocabulary quiz, the teacher still cannot claim the student has no vocabulary.

An example of a case in which ratio data could be used in the behavioral sciences is if a researcher is using as a variable the amount of practicum hours students have logged during a semester. The researcher would give a score of 0 to any student who has seen no clients and logged no hours. Likewise, the researcher could accurately say that a student who has logged 50 hours of client contact during practicum has had twice the amount of client contact as a student who has logged 25 hours.

Although ratio-level measurement is not common in the behavioral and social sciences, its advantages

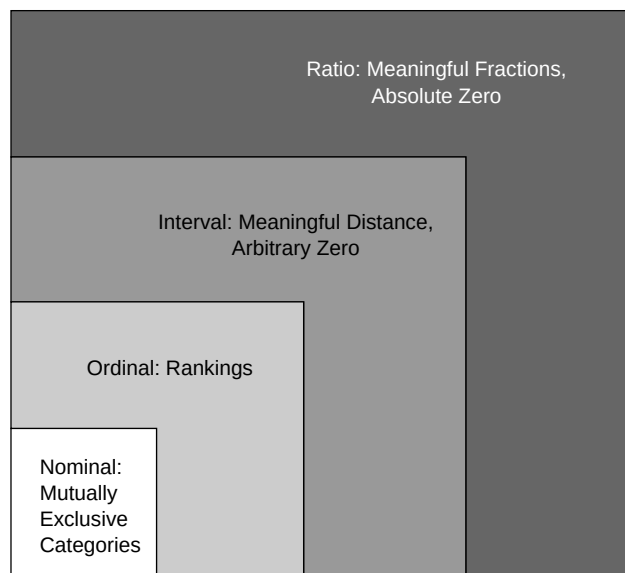


Figure 1 Progressing Levels of Measurement

make it a desirable scale to use. Because it is the most precise level of measurement, and because it contains all of the qualities of the three “lower” levels of measurement, it provides the richest information about the traits it measures.

Ratio data let us know that Person A, who earns an annual income of \$1 million, makes 100 times the amount of money as Person B, who earns an annual income of \$10,000. We can tell more from this comparison than we could if we knew only that Person B makes \$990,000 less per year than Person A (interval-level measurement), that Person A has a larger income than Person B (ordinal-level measurement), or that the incomes of Person A and Person B fall under different socioeconomic categories (nominal-level measurement). This information would give us a more accurate conception of the construct being measured.

—Kristin Rasmussen

See also Interval Level of Measurement; Nominal Level of Measurement; Ordinal Level of Measurement

Further Reading

- Lane, D. (2003). *Levels of measurement*. Retrieved from <http://cnx.rice.edu/content/m10809/latest/>
- Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

Sirkin, R. M. (Ed.). (2005). *Statistics for the social sciences*. Thousand Oaks, CA: Sage.

Trochim, W. M. K. (2002). *Levels of measurement*. Retrieved from <http://www.socialresearchmethods.net/kb/measlevl.htm>

RAVEN'S PROGRESSIVE MATRICES

Raven's Progressive Matrices, published by Lewis, is a test that was originally introduced in 1938. Its purpose is to measure Spearman's *g* factor or general intelligence. For this reason, Raven chose a spatial format for the test, which required the education of relations among abstract items, that is, the visualizing of relationships that are based on perceptual similarities between stimuli.

It is a nonverbal test of reasoning ability that comes in three forms; Coloured Progressive Matrices, Standard Progressive Matrices, and Advanced Progressive Matrices (Sets I and II). The test can be administered in groups or individually, and each form takes from 15 to 30 minutes to complete.

The Coloured Progressive Matrices is a 36-item test that can be used mainly with children from 5 to 11 years of age. In this form, the figures are in color in order to appeal to the children and sustain their attention. The Standard Progressive Matrices consists of 60 items that are presented in five sets, with 12 items per set. The test is employed mostly with persons from 6 to 14 years of age. The Advanced Progressive Matrices is appropriate for older adolescents and adults, primarily for individuals with higher than average intellectual ability. It contains 12 items in Set I and 36 in Set II.

In each form of the test, the examinee is presented with a design or matrix from which a part has been removed. The testee has to look through the matrix and decide which of six or eight pieces given below the matrix is the right one to complete the matrix. The first item in a set is intended to be self-evident, and it is followed by another 11 items of increasing difficulty.

The items referred to are (a) consecutive patterns, (b) analogies between pairs of figures, (c) progressive modifications of patterns, (d) transformations of figures, and (e) disentangling of figures into component parts. Thus, the test estimates the ability to make comparisons, to

deduce by analogy, and to organize spatial patterns into wholes.

The test items can be solved by employing problem-solving strategies. The principle that has to be used to solve each item is either to put forward orally or to draw information from visual perceptual recognition of the stimulus. In the first occasion, an analytical method is used, whereas in the second, a Gestalt procedure involving visual perception is used to solve the items.

The reliability and validity of the test is excellent, but the construct validity of the test is not supported by research because the very easy and very difficult items of the test measure different operations.

Although the test does not respond to the initial Raven's intentions of measuring Spearman's g factor, it is a useful indicator of nonverbal reasoning ability. More specifically, it is easy to administer and useful for screening the intelligence of children and adults with language difficulties and auditory or physical disabilities. Furthermore, it is useful for testees who do not speak English or have a restricted command of English.

—Demetrios S. Alexopoulos

Further Reading

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Sattler, J. M. (1992). *Assessment of children: Revised and updated* (3rd ed.). San Diego, CA: Author.

Raven Standard Progressive Matrices: <http://www.cps.nova.edu/~cpphelp/RSPM.html>

RECORD LINKAGE

Record linkage, or exact matching, refers to the activity of linking together two or more databases on a single population. The U.S. Bureau of the Census uses record linkage in its efforts to estimate the population undercount of the decennial census. The two files that Census links together are a sample of the decennial census and a second, independent enumeration of the population areas covered by the sample. Some individuals are counted in both the census and the second enumeration, whereas others are absent from one or both of the

Table 1 Counts From Two Enumerations Based on Record Linkage

		Second Enumeration		
		Yes	No	Total
Census Enumeration	Yes	n_{yy}	n_{yn}	n_{census}
	No	n_{ny}	?	?
	Total	n_{second}	?	?

canvasses. Suppose that the numbers of individuals who are enumerated are given in Table 1. The question marks indicate counts of individuals that are not known.

The total size of the population can be estimated if assumptions about the two enumeration efforts and the population are made. Under standard assumptions of capture-recapture models, the total size of the population can be estimated as $n_{\text{census}}n_{\text{second}}/n_{yy}$. If 250 people were counted in the census sample, 200 were counted in the second enumeration, and 125 were common to both lists, then the population size would be estimated as $250(200)/125 = 400$. However, if only 100 people were common to both lists, then one would estimate the population size to be $250(200)/100 = 500$.

Record linkage is challenging when the sizes of the files being linked are very large and unique identifying information on every individual is not available. Examples of unique identifiers (IDs) include Social Security numbers (SSNs); U.S. passport numbers; state driver's license numbers; and, except for identical twins, a person's genetic code. The decennial census does not collect SSNs or any other unique ID number. The number of people in the census undercount sample is a few hundred thousand. Thus, record linkage in this context needs to be computerized and automated.

Entries in the two databases are compared on the fields of information common to two files. Consider the following hypothetical records in the two files called File A and File B:

File A Record

Wayne Feller
Male, Married, Age 70
202 Snedecor Rd.
Ames, Iowa

File B Record

W. A. Fuller
Male, Married, Age 71
202 Snedecor, Apt. 3
Aimes, IA

These records, although containing clear differences, could correspond to the same person. Alternate versions of names and addresses, nicknames and abbreviations, and misspellings and typographical errors are frequently encountered in large, population databases. The U.S. Bureau of the Census and other U.S. and foreign statistical agencies use sophisticated methods to address these and other challenges.

Large-scale record linkage operations typically standardize information such as names and addresses, create groups or blocks of records by geography and other factors, and score the degree of agreement between pairs of potential matches. Scores can be based on rules for adding points, also called weights, for agreements and subtracting points for disagreements. A pair of records with a high score is likely to correspond to a single person, whereas a pair with a low score is likely to represent two different people. The weights can be based on characteristics of records in the two files or estimated using statistical models, such as latent class models. The accumulation of similarities and dissimilarities, such as those found when comparing the two hypothetical records above, lead one to believe that a pair of records represents either one person or two different people.

Record linkage, or exact matching, is different from statistical matching. A use of statistical matching is to create matched pairs of people who are similar to one another on key variables, such as age, sex, race, and health status. Matched pairs can be used, for example, in experiments and observational studies to compare medical treatments.

Record linkage (RL) is used in many applications. Besides population undercount estimation, Census and major survey organizations use RL in the creation of sample frames of addresses, businesses, and establishments. The National Center for Health Statistics uses RL for comparing employment records and historical survey records to the National Death Index, a compilation of state death certificate records. Health insurance and medical organizations can use record linkage to link patient records over time in order to form cumulative health histories. Record linkage in these and other applications enable

statistical analyses and studies that otherwise would not be possible. Of course, there is a great concern for the confidentiality of respondent information and the privacy of individuals that must be protected when RL is used.

—Michael D. Larsen

See also Latent Class Analysis; Mixture Models

Further Reading

- Alvey, W., & Jamerson, B. (1997). *Record linkage techniques—1997, Proceedings of an International Workshop and Exposition*. Federal Committee on Statistical Methodology, Office of Management and Budget. Retrieved from http://www.fcsm.gov/working-papers/RLT_1997.html
- Belin, T. R., Ishwaran, H., Duan, N., Berry, S., & Kanouse, D. (2004). Identifying likely duplicates by record linkage in a survey of prostitutes. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. New York: Wiley.
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183–1210.
- Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14, 491–498.
- Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222–230.
- Larsen, M. D., & Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32–41.
- Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford, UK: Oxford University Press.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, pp. 954–959.
- Winkler, W. E. (1994). Advanced methods for record linkage. *1994 Proceedings of the American Statistical Association*, Survey Research Methods Section (pp. 467–472). Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/1994_077.pdf
- Winkler, W. E. (2000). Machine learning, information retrieval, and record linkage. *2000 Proceedings of the American Statistical Association*, Survey Research Methods Section (pp. 20–29). Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/2000_003.pdf

REGRESSION ANALYSIS

Regression analysis is the name for a family of techniques that attempts to predict one variable (an outcome or dependent variable) from another variable, or set of variables (the predictor or independent variables).

We will illustrate this first with an example of linear regression, also called (ordinary) least squares (OLS) regression. When people say “regression” without any further description, they are almost always talking about OLS regression. Figure 1 shows a scatterplot of data from a group of British ex-miners, who were claiming compensation for industrial injury. The x-axis shows the age of the claimant, and the y-axis shows the grip strength, as measured by a dynamometer (this measures how hard the person can squeeze two bars together).

Running through the points is the *line of best fit*, or *regression line*. This line allows us to predict the conditional mean of the grip strength—that is, the mean value that would be expected for a person of any age.

The line of best fit, or regression line, is calculated using the least squares method. To illustrate the least squares method, consider Figure 2, which is simplified, in that it has only four points on the scatterplot. For each point, we calculate (or measure) the vertical distance between the point and the regression

line—this is the residual, or error, for that point. Each of these errors is squared and these values are summed. The line of best fit is placed where it minimizes this sum of squared errors (or residuals)—hence, it is the least squares line of best fit, which is sometimes known as the ordinary least squares line of best fit (because there are other kinds of least squares lines, such as generalized least squares and weighted least squares). Thus, we can think of the regression line as minimizing the error (note that in statistics, the term *error* is used to mean deviation or wandering, not mistake).

The position of a line on a graph is given by two values—the height of the line and the gradient of the line. In regression analysis, the gradient may be referred to as b_1 or β_1 (β is the Greek letter beta). Of course, because the line slopes, the height varies along its length. The height of the line is given at the point where the value of the x-axis (that is, the predictor variable) is equal to zero. The height of the line is called the intercept, or y-intercept, the constant, b_0 (or β_0), or sometimes α (the Greek letter alpha).

Calculation of the regression line is straightforward, given the correlation between the measures. The slope of the line (b_1) is given by

$$b_1 = \frac{r \times s_y}{s_x},$$

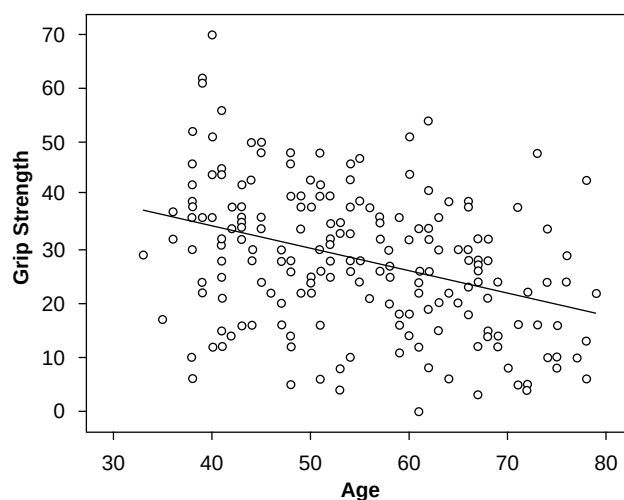


Figure 1 Scatterplot Showing Age Against Grip Strength With Line of Best Fit

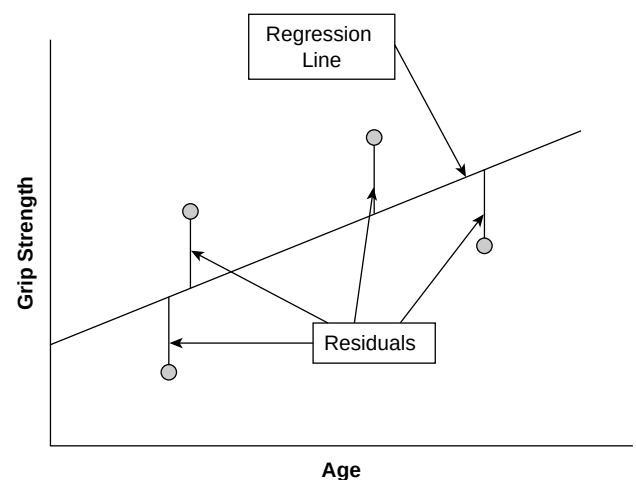


Figure 2 Example of Calculation of Residuals

where

r is the correlation between the two measures,
 s_y is the standard deviation of the outcome variable, and
 s_x is the standard deviation of the predictor variable.

The intercept is given by

$$b_0 = \bar{y} - b_1 \bar{x},$$

where

i is the intercept,
 $\bar{y}$ is the mean of the outcome variable, and
 $\bar{x}$ is the mean of the predictor variable.

In the case of the data shown in Figure 1, the intercept is equal to 50.9, and the slope is -0.41 . We can calculate the predicted (conditional mean) grip strength of a person at any age, using the equation

$$\hat{s} = 50.9 + a \times -0.41,$$

where $\hat{s}$ is the predicted strength and a is the age of the individual. Notice the hat on top of the s , which means that it is *predicted*, not actual. A very similar way to write the equation would be

$$\begin{aligned}\hat{s} &= 50.9 + a \times -0.41 + e \\ s &= \hat{s} + e.\end{aligned}$$

In this equation, we are saying that s is the person's actual strength, which is equal to the expected value plus a deviation for that individual. Now we no longer have the hat on the s , because the equation is stating that the person's *actual* score is equal to that calculated, plus e , that is, error.

Each of the parameters in the regression analysis can have a standard error associated with it, and hence a confidence interval and p value can be calculated for each parameter.

Regression generalizes to a case with multiple predictor variables, referred to as multiple regression. In this case, the calculations are more complex, but the principle is the same—we try to find values for the

parameters for the intercept and slope(s) such that the amount of error is minimized. The great advantage and power of multiple regression is that it enables us to estimate the effect of each variable, controlling for the other variables. That is, it estimates what the slope would be if all other variables were controlled.

We can think of regression in a more general sense as being an attempt to develop a *model* that best represents our data. This means that regression can generalize in a number of different ways.

Types of Regression

For linear regression as we have described it to be appropriate, it is necessary for the outcome (dependent) variable to be continuous and the predictor (independent) variable to be continuous or binary. It is frequently the case that the outcome variable, in particular, does not match this assumption, in which case a different type of regression is used.

Categorical Outcomes

Where the outcome is binary—that is, yes or no—logistic or probit regression is used. We cannot estimate the conditional mean of a yes/no response, because the answer must be either yes or no—if the predicted outcome score is 0.34, this does not make sense; instead, we say that the probability of the individual saying yes is 0.46 (or whatever it is). Logistic or probit regression can be extended in two different ways: For categorical outcomes, where the outcome has more than two categories, multinomial logistic regression is used. Where the outcome is ordinal, ordinal logistic regression is used (SPSS refers to this as PLUM – PoLytomous Universal Models) and models the conditional likelihood of a range of events occurring.

Count Outcomes

Where data are counts of the number of times an event occurred (for example, number of cigarettes smoked, number of times arrested), the data tend to be positively skewed, and additionally, it is only sensible to predict an integer outcome—it is not possible to be arrested 0.3 times, for example. For count outcomes

of this type, Poisson regression is used. This is similar to the approaches for categorical data, in that the probability of each potential value is modeled—for example, the probability of having been arrested 0 times is 0.70, one time is 0.20, three times 0.08, and four times 0.02.

Censored Outcomes

Some variables are, in effect, a mixture of a categorical and a continuous variable, and these are called censored variables. Frequently, they are cut off at zero. For example, the income an individual receives from criminal activities is likely to be zero, hence it might be considered binary—it is either zero or not. However, if it is not zero, we would like to model how high it is. In this situation, we use Tobit regression—named for its developer, James Tobin, because it is Tobin’s Probit regression (Tobin himself did not call it this, but it worked so well that it stuck). Another type of censoring is common where the outcome is time to an event, for example, how long did the participant take to solve the problem, how long did the patient survive, or how long did the piece of equipment last. Censoring occurs in this case because, for some reason, we didn’t observe the event in which we were interested—the participant may have given up on the problem before he or she solved it, the patient may have outlived the investigator, or the piece of equipment may have been destroyed in a fire. In these cases, we use a technique called Cox proportional hazards regression (or often simply Cox regression).

Uses of Regression

Regression analysis has three main purposes: prediction, explanation, and control.

Prediction

A great deal of controversy arises when people confuse the relationship between prediction in regression and explanation. A regression equation can be used to predict an individual’s score on the outcome variable of interest. For example, it may be the case that students who spend more time drinking in bars perform less well in their exams. If we meet a student

and find out that he or she never set foot inside a bar, we might predict that he or she will be likely do better than average in his or her assessment. This would be an appropriate conclusion to draw.

Explanation

The second use of regression is to *explain* why certain events occurred, based on their relationship. Prediction requires going beyond the data—we can say that students who spend more time in bars achieve lower grades, but we cannot say that this is *because* they spend more time in bars. It may be that those students do not like to work, and if they didn’t spend time in bars, they would not spend it working—they would spend it doing something else unproductive. Richard Berk has suggested that we give regression analysis three cheers when we want to use it for description, but only one cheer for causal inference.

Control

The final use of regression is as a control for other variables. In this case, we are particularly interested in the residuals from the regression analysis. When we place the regression line on a graph, everyone above the line is doing better than we would have expected, given his or her levels of predictor variables. Everyone below the line is doing worse than we would have expected, given his or her levels of predictor variables. By comparing people’s residuals, we are making a fairer comparison. Figure 2 reproduces Figure 1, but two cases are highlighted. The case on the left is approximately 40 years old, the case on the right approximately 70 years old. The 40-year-old has a higher grip strength than the 70-year-old (approximately 31 vs. approximately 28 kg). However, if we take age into account, we might say that the 40-year-old has a lower grip strength than we would expect for someone of that age, and the 70-year-old has a higher grip strength. Controlling for age, therefore, the 70-year-old has a higher grip strength.

We will give a concrete example of the way that this is used. The first example is in hospitals in the United Kingdom. Dr Foster (an organization, rather than an individual) collates data on the quality of care in different hospitals—one of the most important

variables that it uses is the standardized mortality ratio (SMR). The SMR models an individual's chance of dying in each hospital. Of course, it would be unfair to simply look at the proportion of people treated in each hospital who died, because hospitals differ. They specialize in different things, so a hospital that specializes in heart surgery would have more patients die than a hospital that specialized in leg surgery. Second, hospitals have different people living near them. A town that is popular with retirees will probably have higher mortality than a hospital in a town that is popular with younger working people. Dr Foster attempts to control for each of the factors that is important in predicting hospital mortality, and then calculates the standardized mortality ratio, adjusting for the other factors. It does this by carrying out regression, and examining the residuals.

—Jeremy Miles

Further Reading

- Berk, R. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.
- Cohen, J., Cohen, P., Aiken, L., & West, S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Robbins, J. M., Webb, D. A., & Sciamanna, C. N. (2004). Cardiovascular comorbidities among public health clinic patients with diabetes: The Urban Diabetics Study. *BMC Public Health*, 5, 15. Retrieved from <http://www.biomed-central.com/1471-2458/5/15>

RELATIVE RISK

The relative risk (RR) is the most common measure epidemiologists use to quantify the association between a disease and a risk factor. It may be most easily explained in the case where both the disease and the risk factor may be viewed as binary variables. Let p_1 and p_0 denote the probabilities that a person with the risk factor develops the disease and that an unexposed person develops the disease, respectively. Then the relative risk is given by $RR = p_1/p_0$. In the

following discussion, we will take risk factors to be exposures. These could be toxic substances to which a subject is exposed, such as cigarette smoke or asbestos, or potentially beneficial foods, such as cruciferous vegetables. But risk factors can also refer to the expression of the disease in relatives or to genetic polymorphisms, to name a few examples. Regardless of the disease or risk factor, an RR greater than one indicates an association between the risk factor and development of the disease, whereas an RR less than one indicates that the risk factor has a protective effect. An important feature of the RR is its multiplicative interpretation—if the RR for a particular exposure is 5, then an individual with that exposure is five times more likely to develop the disease than an unexposed individual.

The most direct way to estimate the relative risk is through a prospective study. In such a study, exposed and unexposed individuals who are free of the disease are recruited into the study and then followed for disease information. After a period of time long enough for sufficient numbers of the patients to develop the disease, the relative risk may be calculated as the ratio of the proportion of exposed subjects who develop the disease to the proportion of unexposed subjects who develop it. A well-known example is a study of lung cancer and smoking among British doctors. The study, reported in the *British Medical Journal* by Doll and Peto in 1976, with a follow-up in 1994, involved recruiting 34,440 British doctors in 1951. The subjects filled out questionnaires on smoking habits and were followed for 40 years. The risks of death from lung cancer among smokers and lifelong nonsmokers were, respectively, 209 and 14 per 100,000. The relative risk is thus 14.9. As this example illustrates, prospective studies require large numbers of subjects who are followed for long periods of time, often decades. Such studies are often the most definitive in establishing a causal relationship, and they allow one to study the effects of the exposure on a variety of diseases. But the time and expense of conducting them can be prohibitive. Furthermore, for very rare diseases, it may not be possible to recruit enough healthy subjects to produce

enough disease cases to calculate a reliable estimate of the RR.

Case control (or retrospective) studies are an alternative form of study that can be used to estimate the RR for an exposure and a rare disease. In a case control study, the sampling is the reverse of that in a prospective study. One starts by sampling a set of individuals who have already developed the disease and a set of control patients who have similar characteristics to the cases, but who haven't developed the disease. One then compares the proportion of cases exposed to the risk factor to the proportion of controls who were exposed.

To understand how a case control study can estimate the RR requires an understanding of two key concepts. The first is the odds ratio (OR), which we will explain in terms of a prospective study. The odds of developing the disease among exposed individuals are $p_1/(1 - p_1)$, and the odds among unexposed individuals are $p_0/(1 - p_0)$. Then the odds ratio is given by the ratio of these two quantities, which may be expressed as $OR = p_1(1 - p_0)/p_0(1 - p_1)$. If the disease is rare, so that both p_1 and p_0 are small, then the OR is a good approximation of the RR.

The second key concept underlying case control studies is that the odds ratio of *exposures* estimated for cases and controls from a case control study equals the odds ratio of *disease* for exposed and unexposed individuals that would be estimated from a prospective study. To see this, let P_1 denote the probability of exposure to the risk factor among the cases and P_0 the probability of exposure among the controls. Then it may be shown that

$$OR = \frac{p_1(1 - q_0)}{q_1(1 - p_0)} = \frac{P_1(1 - Q_0)}{Q_1(1 - P_0)}.$$

That is, the odds ratio for exposure for cases versus controls, computed from a case control study, is equal to the odds ratio for disease for exposed versus nonexposed individuals from a prospective study. Thus, for a rare disease, the odds ratio computed from a case control study is a good approximation to the relative risk.

Case control studies require far fewer individuals than do prospective studies, and they can be carried out in a much shorter period of time. For these reasons, case control studies are the most common observational tool epidemiologists use for measuring relative risks. Bear in mind that the validity of case control studies depends on the assumption that the sampling of cases and controls are random samples from the population of interest. Thus, identifying and adjusting for potential confounding factors is essential to a successful case control study.

Logistic regression is the main statistical tool for estimating odds ratios that adjust for risk factors. The binary outcome is disease status, and the primary predictor is exposure. The exposure variable may be binary, as discussed above, or it could be a quantitative measure of the level of exposure. Other covariates may be included in the model to adjust for other factors. The parameter estimate associated with exposure is the log-odds ratio, which may be interpreted as the log relative risk for rare diseases. A related form of epidemiological study that one may use to estimate the RR is the *matched* case control study. A simple version of this uses multiple sets of one case and one control each, matched for potential confounding factors. A common variation is to create sets with one case and two to five controls each. With matched designs, a modified form of logistic regression known as *conditional* logistic regression is necessary to obtain unbiased estimates of the RR. The suggested readings contain more details about these methodologies.

—Dirk F. Moore

Further Reading

- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research, Volume I. The analysis of case-control studies*. Lyon, France: International Agency for Research on Cancer.
- Doll, R., Peto, R., Wheatley, K., Gray, R., & Sutherland, I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal*, 309(6959), 901–911.
- Jewell, N. P. (2004). *Statistics for epidemiology*. New York: Chapman & Hall/CRC.

RELIABILITY THEORY

Reliability and validity are two major requirements for any measurement. Validity pertains to the correctness of the measure; a valid tool measures what it is supposed to measure. Reliability pertains to the consistency of the tool across different contexts. As a rule, an instrument's validity cannot exceed its reliability, although it is common to find reliable tools that have little validity.

There are three primary aspects to reliability: (a) A reliable tool will give similar results when applied by different users (such as technicians or psychologists). (b) It will also yield similar results when measuring the same object (or person) at different times. In psychometrics, reliability also implies a third feature, which is relevant to scales (measures that include various submeasures or items). Specifically it entails the requirement that (c) all parts of the instrument be interrelated.

1. *Interrater Reliability*: Some measurements require expertise and professional judgments in their use. The reliability of such a tool is contingent on the degree to which measurements of the same phenomenon by different professionals will yield identical results. What we want to avoid is a test that essentially relies on undefined judgments of the examiner, without concrete criteria that are clearly spelled out.

Statistically, this aspect of reliability is usually determined by having several raters measure the same phenomena, and then computing the correlations between the different raters. For typical measures that yield numerical data, a correlation index needs to be high (e.g., in the .90s) to demonstrate good interrater reliability.

2. *Test-Retest Reliability*: This feature, often referred to as *temporal stability*, reflects the expectancy that the measurement of a specific object will yield similar results when it is measured at different times. Clearly, this is based on an assumption that one does not expect the object to be changing between the

two measurements. In fact, this may not be the case for many constructs. (Consider, for example, blood pressure or stress, both of which would be expected to vary from one time to another—even for the same person.) Specifically within psychology, it is important to understand that this aspect of reliability pertains only to measures that refer to traits (aspects of personality that are constant regardless of environmental events or context); it does not pertain to states (specific aspects of behavior or attitudes that vary based on situations and interactions at the moment).

Statistically, this aspect of reliability is usually determined by having a group of people measured twice with the instrument. The time interval can vary, based on specific studies, from several weeks to a year or two. Unfortunately, testing experts often choose a short duration between tests (there are published test-retest time periods of only 6 hours!), which make their claim of testing an actual trait equivocal. Correlations are computed between the two trials. For typical measures that yield numerical data, a correlation index needs to be high to demonstrate good interrater reliability (although .70 would be sufficient).

3. *Internal Consistency*: This usually entails the reliability of measures that have multiple items; such measures are known as scales.

Consider, for example, a scale of irritability. Such a scale might include 20 items that focus on different aspects of this construct. One might ask how often the person gets into arguments, another might look into how often the person changes doctors, whereas a third could ask about the person's rejection patterns of potential dates. Obviously, the experts who constructed this scale believed that all of these items are indicative of irritability. Indeed, it may well be true that irritable people would show high scores on all of these items. However, it is feasible that the major reason most people reject dates differs significantly from the reason people change doctors.

Another aspect of the problem of internal consistency might be understood from the perspective of the meaning of the score on a scale. Take, for example, the case of a firm that is interested in hiring interviewers

who will be pleasant to potential customers. Such a firm might decide to use a 100-item scale of pleasantness, which contains imagined scenarios that are presented to respondents, each having a specific response that is considered “pleasant.” In the real world, the firm cannot expect to find sufficient interviews that will score 100, so the firm decides to hire those candidates who score highest. Say that two candidates scored 98; one of the respondents “failed” two items indicating that she would ask loud people to keep it down in a theater and that she would not allow someone to cut in line in front of her at a bank. The other respondent “failed” two items indicating that she has been a party to a lawsuit and that she would hang up on a telemarketer. By assigning these respondents equal scores, the instrument implies that they are at an equivalent level of pleasantness. But what evidence is there that the items they failed are just as meaningful in terms of the overall scale? Furthermore, is it possible that there are different factors in unpleasantness, and that these items don’t really speak to the same construct?

There are several methods of establishing internal consistency, all of which are based on intercorrelations between items. One method is called split half, where the scale is divided into different sets, each containing half the items, and the halves are then tested for an acceptable relationship using the correlation statistic. Often, item-total correlations are used to establish internal consistency (that is, correlations between each of the items and the total scale score). Typically, an item must correlate .3 or higher with the total score in order to remain part of the scale. A popular overall statistic that takes into account all of the possible item intercorrelations is Cronbach’s alpha (alternatively, the Kuder Richardson-20 formula for dichotomous items), where an alpha of .70 or higher is considered acceptable to establish this form of reliability.

Scaling issues that are related to internal consistency often are found in weighting procedures. Some tests contain items that are given more weight in the total score than others. These features introduce statistical complications in the process of establishing consistency, which are often solved through multiple regression methods.

All in all, reliability is a major part of the preparatory work that goes into scale construction, often in conjunction with factor analysis. It is not featured emphatically in test descriptions, but poor reliability will doom a test’s validity and its usefulness to measure anything meaningful.

—Samuel Juni

Further Reading

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78*, 98–104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85–98.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.
- Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358–376.

REPEATED MEASURES ANALYSIS OF VARIANCE

In a repeated measures analysis of variance, we are faced with the task of comparing means of groups that are dependent. Unlike the usual analysis of variance (ANOVA), where the groups are independent, in repeated measures ANOVA, the groups and the group means are dependent. Because the group means are dependent, we must adjust the usual statistical and inferential processes to take the dependencies into account. Before going on with our discussion of repeated measures ANOVAs, let’s consider two situations that are likely to yield correlated means.

To begin with, consider a situation in which we have 10 raters (judges) and five job applicants. In this situation, each rater would rate each candidate, but the five ratings generated by a rater are very likely to be dependent. If we were to create a table with candidates making the columns and raters making the rows, then the observations in each column would be independent, but not so across each row. The mean ratings for the candidates would then be dependent; comparing the means requires a procedure that takes into consideration the dependencies.

Consider another example where individuals are observed over time. In this type of experimental design, one or more groups of individuals are observed before and after a treatment. This design is traditionally called a pre-post design, and the number of posttreatment observations is usually greater than one. For instance, the individuals can be observed at 3-month intervals for a period of a year, yielding four posttreatment observations and one pretreatment observation. Just as in the previous example, these five means are likely to be correlated, and comparing them necessitates a procedure that takes the correlation into account.

We have seen two examples in which the means are probably dependent. However, there are many additional situations that can generate dependent means. For instance, using groups that have been matched along specific criteria to make them equivalent can also undermine the assumption of independence. Researchers must be vigilant and always look for design characteristics or constraints that may render the means dependent.

When repeated measures designs are compared to other designs, we see a number of advantages and disadvantages. Some of the advantages are that (a) they use fewer subjects; (b) they tend to be more efficient (subject variability is controlled better, which often increases power); and (c) they cost less to implement. Of course, not all studies can use repeated measures. There are many situations that preclude the use of repeated measures. For instance, if a prior treatment affects performance on a subsequent treatment—due to contrast effects, fatigue, emotional reaction, and so on—and one can't control for this nuisance effect, a

repeated measures design would not be appropriate. (Keppel and Wickens have provided more information on nonstatistical problems that may affect the interpretation and execution of repeated measures studies.)

These advantages do not come without disadvantages. The major disadvantages of repeated measures designs are that they are harder to analyze and interpret. They are harder to analyze because the groups are not independent, and they are harder to interpret and design because of the possibility of carryover or similar effects.

Traditionally, the data in most analyses of variance (ANOVAs) have been analyzed with the F test. The assumptions for the F test are normality, homogeneity of variances, and independence. As it works out, however, the F test is capable of handling certain dependencies between groups. These dependencies are subsumed under what is called the circularity or sphericity assumption. The repeated factors and the interactions involving repeated factors require the circularity assumption for the validity of the F ratios. (The assumption in essence transfers the independence requirement to the contrasts implicitly tested by the F test.) Thus, the assumptions for the repeated factors are normality and circularity.

The most restrictive form of the circularity assumption demands that the observations between any two levels of the repeated factor be equally correlated. In a simple repeated measures design with three levels in the repeated factor T , this restrictive form of the assumption requires that the three correlations between the levels of the repeated factor be equal:

$$r(T_1, T_2) = r(T_1, T_3) = r(T_2, T_3).$$

This extreme form of the circularity assumption is known as the compound symmetry assumption. If the circularity assumption is met, the usual F tests are valid and the repeated measures analysis resembles the usual ANOVA procedure.

It is not unusual, however, for the circularity assumption to fail. In these situations, we must reduce the degrees of freedom (df) of the F test before looking up the critical value in the F table. Alternatively,

we could have our computer package find the p value for the F ratio with the reduced degrees of freedom. The degrees of freedom are reduced to the degree that we violate the circularity assumption. More severe violations reduce the df more than less severe violations. The F tests based on reduced df are approximate in that they depend on the unknown population variances and covariances between the levels of the repeated factors, which are never known in practice. However, it has been found that procedures are very accurate when they are based on sample estimates of these parameters.

Many computer packages, in addition to providing tests that reduce the df , provide multivariate tests of the univariate repeated measures hypotheses. If one conceptualizes a repeated measures design as a multivariate analysis of variance (MANOVA), we can test the univariate hypotheses by creating a set of dependent variables that represents the differences between the levels of the repeated factor. By conceptualizing a repeated measures design as a multivariate analysis of variance, we get around the circularity assumption. The MANOVA makes no structural assumption about the variance-covariance matrix.

With the new advances in the analysis of mixed model designs, we can yet provide another approach to the analysis of repeated measures data. We can estimate the structure of the variance-covariance matrix underlying the repeated measures and carry out a large-sample maximum likelihood analysis of the repeated factor; such approach can be found in SAS's Proc Mixed and in other computer packages. The maximum likelihood approach is generally more efficient than the multivariate approach when we have missing data.

All of these approaches to the analysis of repeated measures have different strengths and weaknesses. The choice of what technique to use depends on the assumptions that one can make, the availability and familiarity with computer packages, and the researcher's preference. To this date, no clear winner has been identified. However, it is clear that routinely we should carry out more than one procedure and never assume that we have met the circularity assumption. Next, we provide an example in which the circularity assumption is met, because there are

only two levels of the repeated factor. Although the example is computationally simple in that we do not need to be concerned with circularity, it illustrates the fundamental ideas behind the analysis of a repeated measures design.

Example

Does flexibility training help improve spinal extension (bending backward) in elderly women? Based loosely on research by Rider and Daly, this research had 20 female volunteers with an average age of 71.8, who were in moderately good health including no orthopedic problems. Ten women were randomly assigned to the experimental condition, a 10-week program with 3 days per week of spinal mobility and flexibility exercises. The other 10 women were randomly assigned to a control condition, which was a regular exercise program without flexibility training. All women were measured before and after the 10-week program. Table 1 shows the scores on a spinal extension measure (higher scores mean better extension).

Table 1 Spinal Extension Scores

		<i>Pretest</i>	<i>Posttest</i>
<i>Experimental</i>	S1	13	26
	S2	19	27
	S3	21	28
	S4	15	26
	S5	12	17
	S6	26	30
	S7	25	29
	S8	11	15
	S9	16	25
	S10	21	27
<i>Control</i>	S11	15	19
	S12	18	20
	S13	26	26
	S14	23	23
	S15	24	26
	S16	19	19
	S17	14	16
	S18	20	20
	S19	11	9
	S20	20	25

Table 2 The ANOVA Procedure

Dependent Variable: spinalext

Tests of Hypotheses Using the ANOVA MS for $s(g)$ as an Error Term

Source	DF	ANOVA SS	Mean Square	F Value	Pr > F
g	1	32.40000000	32.40000000	0.68	0.4201

Tests of Hypotheses Using the ANOVA MS for $t*s(g)$ as the Error Term

Source	DF	ANOVA SS	Mean Square	F Value	Pr > F
t	1	176.4000000	176.4000000	49.23	<.0001
$g*t$	1	84.1000000	84.1000000	23.47	0.0001

Table 2 shows the F tests from an SAS ANOVA on these data, with F ratios for g (experimental vs. control), t (pretest vs. posttest), and the $g \times t$ interaction.

The results show that trial effect and the group-by-trial interaction are significant. There was no group effect. The significant F ratio for the group-by-trial interaction shows that spinal extension flexibility differences between the pretest and posttest depend upon whether the women were in the experimental group or control group. Using multiple comparison procedures, we can explain this interaction by showing a significant gain in flexibility from pretest to posttest for the experimental group but not for the control group.

This example illustrates that in pre–posttest designs, the hypothesis of interest is often the interaction hypothesis and not the main effects. In this example, we met the circularity assumption. Had we presented an example in which the circularity assumption had not been met, we would have proceeded in the same manner except that we would have evaluated the interaction and trials effects with corrected F ratios (or multivariate tests).

—Jorge L. Mendoza and Larry Toothaker

Further Reading

- Collier, R. O., Jr., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios, in repeated measures designs. *Psychometrika*, 32, 339–353.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 86, 964–973.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the two-way repeated measure design. *Multivariate Behavioral Research*, 9, 165–178.

Rider, R. A., & Daly, J. (1991). Effects of flexibility training on enhancing spinal mobility in older women. *Journal of Sports Medicine and Physical Fitness*, 31(2), 213–217.

Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 32, 269–286.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

RESIDUALS

Residuals play an important role in statistical modeling. The most common definition of residual is the difference between the observed and expected response based on the estimated model so that each observation has a corresponding residual. This definition of residual is frequently encountered in most regression models and analysis of variance models. Loosely speaking, residual is interpreted as the portion of the observed response not explained by the model with the set of predictor variables.

Use of Residuals

A key step in modeling is the examination of the residuals for the purpose of checking if the model is a good fit in light of the data. In particular, residuals are used to check the validity of the assumptions of the model and to identify observations that are considered outliers (unusually large or small observations) and influential in estimating the model. These may be achieved through graphical techniques or performing formal statistical tests applied to residuals. One may work with raw residuals or standardized residuals. Standardized residuals are obtained

by dividing the raw residuals by their estimated standard error.

An Illustration

In ordinary regression modeling, a response variable (denoted by Y) is modeled using a linear relationship with a single variable, known as the predictor variable (denoted by X). Mathematically, this model is represented by the following equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are the intercept and slope of the line, respectively, and ε is the random error term. As an example, suppose it is of interest to model job satisfaction (Y) based on the number of years on the job (X). The data are given in Table 1. A scatterplot of the data is displayed in Figure 1, represented by dots. The estimated model is given by

$$\hat{Y} = 4.83 + 0.17 X,$$

which is represented by the solid line in Figure 1. Residual is graphically represented as the difference between a given observation and the fitted line, as illustrated by the vertical line in Figure 1. For this particular observation, $Y = 5.2$ and $X = 7$, so that the estimated response is given by

$$\hat{Y} = 4.83 + (0.17) \cdot (7) = 6.02.$$

Table 2 Data on Job Satisfaction and Number of Years on the Job

<i>Number of Years</i>	<i>Job Satisfaction</i>
7	5.2
4	6
13	7
10	5.9
15	7
12	7.3
9	6.5
14	8.1
20	8

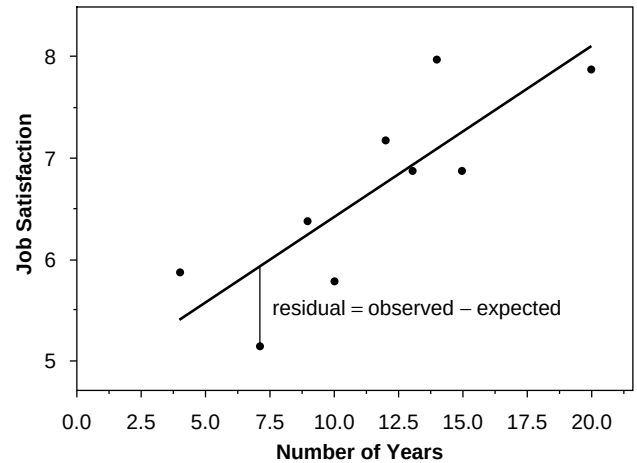


Figure 1 Scatterplot of Job Satisfaction Data With Best Fit Line

Consequently, the residual for this observation is

$$e = Y - \hat{Y} = 5.2 - 6.01 = -0.82.$$

A plot of the residuals versus the predictor variable checks the assumption that the relationship between X and Y is linear. This residual plot applied to the current example is displayed in Figure 2. If the assumption of linearity is appropriate, there should not be any obvious systematic pattern in this residual plot, which is true in this case. Hence, there is no evidence to reject the linearity assumption of the model.

Other Types of Residuals

In more complex statistical models, residuals are defined differently, and, in some cases, there may be

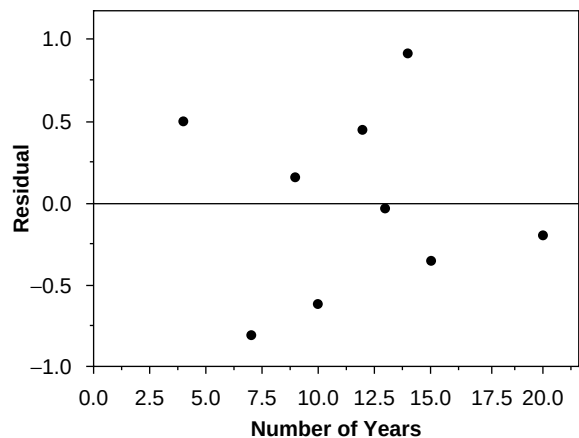


Figure 2 Residuals Versus Number of Years on the Job

more than one type of residual. For instance, in logistic models, there are deviance residuals in addition to the commonly defined residuals. In survival models, where the response variable of interest is time until an event occurs, there are also deviance, Cox-Snell, martingale, and score residuals, to name a few. The definitions of these residuals are no longer as simple as observed minus expected. However, no matter how these residuals are defined, their main purpose is still the same, that is, checking the appropriateness of the model.

—*Inmaculada Aban*

See also Goodness-of-Fit Tests; Regression Analysis

Further Reading

- Cook, R., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.
- Cox, D., & Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, B*, 30, 248–275.
- Kleinbaum, D., Kupper, L., Muller, K., & Nizam, A. (1998). *Applied regression analysis and multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury.
- Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). Chicago: McGraw-Hill Irwin.

RESPONSE TO INTERVENTION

Response to Intervention (RTI) refers to a student-centered assessment and intervention model that has been proposed recently as an alternative approach to the identification of specific learning disabilities (SLDs) under the 2004 reauthorization of Individuals with Disabilities in Education Improvement Act (IDEIA). The IDEIA eliminates the longstanding requirement (i.e., since 1977) that a significant discrepancy between IQ and achievement must be demonstrated to classify a child as a child with an SLD. Although states may still use the IQ-achievement discrepancy model, the U.S. Department of Education, in its proposed IDEIA regulations, is encouraging adoption of the RTI model instead. RTI reflects a reconceptualization of how learning disabilities are assessed and identified by determining whether a child responds to

scientific, research-based interventions as part of the evaluation criteria used to determine if the child has a learning disability, regardless of the existence of an IQ-achievement discrepancy.

Background

The term RTI was first conceptualized in 1982 by leaders in the field of education, who proposed that the validity of a special education classification be judged according to three criteria. The first criterion was whether the quality of the general education program was such that adequate learning might be expected. The second criterion was whether the special education program would lead to academic improvements justifying the classification. The third criterion was whether the assessment process was accurate and meaningful. Under this model, all three criteria must be met for the classification to be considered valid. This new framework required judgments about the quality of instruction in both the general education and special education settings as well as judgments about the way students responded to these learning environments, all based on accurate and meaningful assessments.

The main premise of RTI is that students are identified as having an SLD when their response to effective academic interventions is significantly lower than that of peers. The inference is that students who fail to make adequate progress in response to scientific interventions that are proven successful for most students must have a deficit that requires specialized treatment beyond what general education programs can offer. RTI is sometimes referred to as the problem-solving method because it focuses on applying a problem-solving framework to identify and address student difficulties.

Prior to the RTI conceptualization, SLD was viewed exclusively as a within-child deficit, despite awareness that learning is influenced by the context in which instruction takes place. Instead of making the assumption that the underlying cause of the learning difficulty lies within the child, RTI models recognize that the difficulty may also lie within instruction, within the child, or a combination of the two. Thus,

the new approach focuses on early identification/prevention of learning problems and the contribution of the instructional environment to the child's academic growth. Instead of measuring a student's skills at a single point in time and identifying the student as SLD, assessment and intervention occur on an ongoing basis.

RTI Models

Currently, there are many ways to implement RTI; however, there are some common elements that distinguish a RTI model from other approaches to SLD identification. The core characteristics of RTI include the following:

- Students receive high-quality instruction by their classroom teacher in their general education classrooms while their progress is monitored.
- Those who do not respond receive additional or alternative instruction while their progress continues to be monitored.
- Students who still do not respond may qualify for comprehensive special education evaluation or special education services.

This multiple-step process may vary in terms of the number of levels or tiers in the process; the person(s) responsible for the assessments and interventions; and whether the process is a forerunner to a formal comprehensive evaluation for special education eligibility, or if RTI is itself considered the eligibility evaluation. Several components are necessary to enhance the effectiveness of RTI approaches: (a) ongoing, frequent assessment of student progress, (b) knowledge and skill in implementing evidence-based instruction, (c) a system to screen and track the progress of a large number of students, and (d) systematic assessment of the fidelity or integrity with which the assessments and interventions are implemented.

The Three-Tier RTI Model

Among the most well-conceptualized RTI models is the three-tier model proposed by L. S. Fuchs from Vanderbilt University, also known as “treatment

validity.” Tier I uses classwide assessment to determine whether the overall rate of responsiveness to instruction for the general education environment is generally effective so that adequate student progress is expected. Tier II assessment consists of identifying those students whose *level* of performance and *rate* of improvement are significantly below those of classroom peers. These children are identified as *dually discrepant* because their level of performance and their rate of performance (i.e., slope) fall below the level and rate of classmates. In Tier III, alternative or additional evidence-based interventions are implemented in the general education setting to enhance the quality of education. If intervention fails to promote student growth, comprehensive special education evaluation is considered.

Curriculum-Based Measurement in RTI

Curriculum-based measurement (CBM) is a multiple-probe, brief-duration (e.g., 1 minute) assessment method designed to measure student performance over time to identify students whose level and rate of performance are below those of the reference group. Equal importance is given to skill level (low achievement) and progress (slope). A student who is achieving significantly below level and whose rate of progress is similarly deficient is considered to be at risk. This dual discrepancy of low achievement and low rate of progress becomes the index by which responsiveness to instruction is judged. Students who are dually discrepant undergo additional interventions in the general education classroom while being monitored with CBM “probes” (i.e., student brief, timed samples) to determine if they will respond to the additional instruction. Special education evaluation or placement is considered only if they continue to demonstrate a dual discrepancy after an evidence-based intervention has been tried for some time (e.g., 8 weeks).

RTI Models Address Poor Instruction

RTI models address an often overlooked criterion in the IDEIA regulations—whether the student has

received adequate instruction in general education but failed to benefit from it. Under RTI, students cannot be identified as dually discrepant unless there is indication that most students in the classroom are responding to the curriculum. If the classroom as a whole is not performing at the same level and rate as other classrooms, then a classroom-level intervention geared to improving the academic performance of all students occurs before individual deficits are considered. In this way, RTI proponents maintain that they can rule out poor-quality instruction as a cause of a child's learning difficulty.

On the other hand, if a student responds inadequately to instruction that benefits the majority of students, poor-quality instruction can be ruled out as a feasible explanation for poor academic progress, suggesting instead that a disability is responsible for a child's lack of progress. In other words, the possibility of a neurological deficit (a within-the-child deficit) is not ruled out until the child is unresponsive to quality instruction.

Advantages of RTI Models

RTI models can provide assistance more quickly to a greater number of low-performing students than is possible with IQ-achievement discrepancy models. RTI models emphasize classwide assessment, enabling early identification of struggling students. Because RTI focuses on effective instruction and early intervention, it may benefit all children, including those who enter school with limited literacy and language proficiency who may be at risk for failure, as well as children with disabilities. RTI proponents argue that RTI represents a switch from assessment for eligibility purposes to assessment for instructional purposes, and that RTI ensures that student progress is monitored and instructional interventions are tested. Using an RTI model as part of an overall universal screening may also reduce the reliance on teacher referral, thereby decreasing possible referral bias. Finally, because RTI relies on ongoing data collection, this approach can reduce the influence of measurement error that characterizes assessments administered at a single point in time. A more accurate picture of the student may be

possible because the focus is not only on level of performance but also on growth over time.

Controversial Aspects of RTI

In spite of the increasing popularity and acceptance of RTI, there are many controversial issues that require resolution. Critics assert that children who are identified with an SLD under the RTI-only model may not have a true disability, particularly because no description of cognitive deficits can be identified within the RTI framework, although the IDEIA definition of SLD specifically requires cognitive processing deficits. Opponents of RTI argue that if dual discrepancy is a valid SLD marker, then children identified as SLD should be distinguishable from other groups (e.g., low-achievement children). Yet when RTI has been studied, the distributions of children defined as learning disabled and low achievers overlap substantially, and reading improvement is basically the same for both groups.

Another controversial issue is whether failure to RTI may be attributable to other causes (e.g., mental retardation, emotional disturbance) rather than SLD. Critics also point out that RTI requires arbitrary cut-off points to identify those who fail to RTI. However, the arbitrary cut-off point is a limitation for other approaches for SLD identification as well.

Durability of response to instruction is another controversial area. If a student responds to relatively intensive but short-term instruction treatment, the assumption under RTI is that a disability has been ruled out. Although this may be true for many students, research evidence suggests that the difficulty will reemerge for others when the intensive instruction ceases. Feasibility of evidence-based instruction for the general education classroom is another issue. Interventions that are research-based but not feasible are not likely to be implemented with fidelity, which would undercut the validity of RTI decision making.

Implementing RTI in a way that enables procedural standardization across classrooms, schools, districts, and states is also a challenge for RTI. In addition, research support for RTI is uneven. Adequate research support exists for RTI with reading fluency, particularly

in the early elementary grades. CBM probes for assessing growth have been established in mathematics, spelling, and written expression, but research-based intervention methods for these areas await development. Finally, adequately trained personnel are not yet available on a nationwide basis. A growing number of school districts in Florida, Iowa, Kansas, Minnesota, Ohio, Pennsylvania, South Carolina, and Wisconsin already use RTI to identify students for special education services. Yet most efforts have been implemented on a small scale, and most states lacked personnel trained and skilled in RTI when the IDEIA took effect on July 1, 2005. Perhaps more importantly, RTI implementation requires a paradigm shift for many professionals in conceptualizing assessment and intervention, and resistance to change is to be expected.

In spite of the enormity of the task, and the many accompanying challenges, RTI is at the forefront of a nationwide change in the SLD identification process.

—Romilia Domínguez de Ramírez
and Thomas Kubiszyn

Further Reading

- Kovaleski, J., & Prasse, D. P. (2004). Response to instruction in the identification of learning disabilities: A guide for school teams. *NASP Communiqué*, <http://www.nrld.org/html/research/rti/RTIinfo.pdf>
- Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York: Guilford.
- Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford.
- Vaughn, S., & Fuchs, L. S. (Eds.). (2003). Redefining LD as inadequate response to instruction [Special issue]. *Learning Disabilities Research & Practice*, 18(3).

AIMSweb Response to Intervention in a three-tiered model: http://www.aimsweb.com/products/aimsweb_rti.htm
Response to intervention topical links: <http://www.wested.org/nerrc/rti.htm>

REVERSE SCALING

Scales are sets of items measuring similar constructs on questionnaires. For example, a questionnaire

designed to measure students' satisfaction with a course might contain scales on the instructor, content, course workload, and so on. Generally, scale items are constructed so that participants who have a favorable disposition toward the construct being measured would agree, or strongly agree, with the statements. Examples of such items include the following:

- The instructor's explanation of concepts was clear.
- The instructor responded well to student questions.
- The instructor is competent in his/her area.

Reverse coded items are items phrased in the semantically opposite direction. Examples of reverse coded items are as follows:

- The instructor is impatient with students.
- The instructor did not present concepts clearly.

Reverse scaling is the use of reverse coded items on scales.

Purpose of Reverse Scaling

In his presentation of scaling, Likert recommended constructing scales to balance the item wording and phrase approximately half the items in the reverse. The recommendation to use reverse scaling continues because many psychometricians believe that the inclusion of reverse coded items motivates participants to process items more carefully and prevents negative respondent behaviors such as response set, satisficing, and acquiescence. Response set is the tendency of participants to respond to the set of scale items rather than individual items. For example, a respondent may have a positive impression of an instructor being evaluated and simply respond to the entire set of items positively rather than processing the nuances of each item individually. The inclusion of an item worded in the opposite direction could encourage participants to read and process each item more carefully. Psychologists contend that people prefer to answer in agreement rather than disagreement; therefore, participants may acquiesce or simply agree with an item out of some form of social desirability. Furthermore, participants may also satisfice or agree

with an item because doing so requires minimal cognitive effort. Therefore, reverse coded items are prescribed to force participants to process individual items more carefully and accommodate participants who wish to vary their responses and not always provide the same answer.

Coding and Scoring of Reverse Coded Items

For initial coding purposes, it is best to code reverse coded items in the same way that the traditional items are coded (i.e., *strongly agree* = 5, *agree* = 4, *neutral* = 3, *disagree* = 2, and *strongly disagree* = 1). However, once the data have been entered, the scores for all reverse coded items should be recoded to facilitate consistent interpretation of the data and must be recoded if items are to be combined to form scales. Suppose a scale contained four traditionally worded items and one reverse coded item. If a respondent strongly agreed with the construct measured by the scale, they should respond *strongly agree* to the four traditionally worded items and *strongly disagree* to the reverse coded item. Thus, the raw scale score for the respondent would be 21 (i.e., $5 + 5 + 5 + 5 + 1 = 21$). This score of 21 is misleading because of the score of 1 for the reverse coded item. The score of 21, rather than 25, indicates less than strong agreement with the scale items. Therefore, the scores for all reverse coded items must be recoded. Researchers using any instrument must always search for reverse coded items that might be included in the instrument.

To recode reverse coded items, the numerical values assigned to each category must be reversed (i.e., *strongly agree* = 1, *agree* = 2, *neutral* = 3, *disagree* = 4, and *strongly disagree* = 5). If the data are analyzed with Statistical Package for the Social Sciences (SPSS), then you can either *recode into the same variable* or *recode into a different variable*. If you choose to recode into the same variable, then your original data will be overwritten. Therefore, a safer practice is to recode into a different variable. To begin the recoding process,

- On the top tool bar, click *Transform*.
- From the drop-down menu, click *Recode*.
- From the next drop-down menu, click *Into Different Variable*.

Now, a dialog box should appear to *Recode into Different Variables*. The first steps are to select and rename any variables that need recoding. To do this,

- Select the first variable to be recoded from the variable list on the far right column, and then click the arrow to send the variable into the *Numeric Variable → Output Variable* box.
- Enter the information for the new variable in the *Output Variable* area on the far left of the dialog box.
 - Enter the name for the new variable in the *Name* box (e.g., the original variable name with the letter *r* added to the end to indicate that it has been reverse coded).
 - If you choose, enter the label for the new variable in the *Label* box.
 - Click *Change*.
- Repeat this process for all variables to be recoded with the same codes.
- To enter the new codes, click the button *Old and New Values . . .*

Now, a dialog box should appear to *Recode into Different Variables: Old and New Values*. To recode the values,

- On the left side of the box, in the area marked *Old Value*, enter the initial value in the box labeled *Value*.
- On the right side of the box, in the area marked *New Value*, enter the new value in the box labeled *Value*.

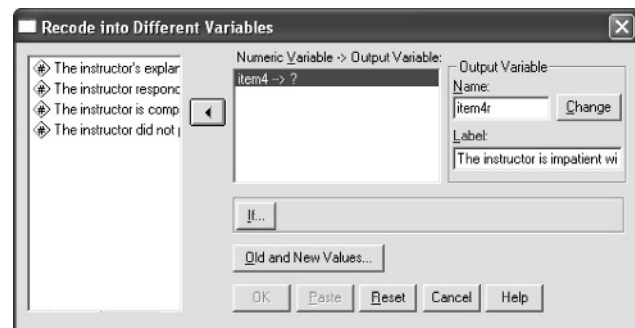


Figure 1 SPSS Recoding Dialog Box

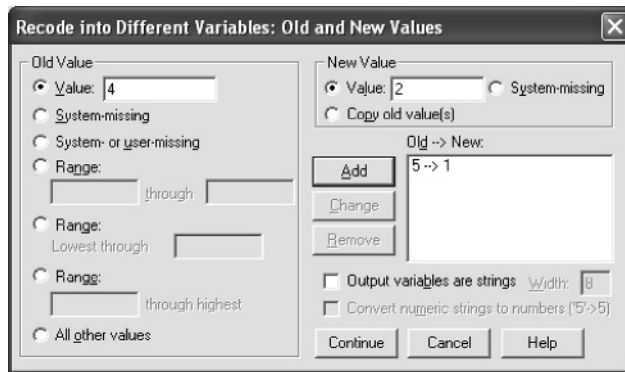


Figure 2 SPSS Old and New Values Dialog Box

- Click *Add*.
- Repeat this process for all values (i.e., $5 = 1$, $4 = 2$, $3 = 3$, $2 = 4$, and $1 = 5$). Check the list that appears in the *Old*→*New* box. This is the only time to view the changes that will be made to the variables.
- Click *Continue*.

The *Recode into Different Variables* should reappear. To complete the recoding, click *OK*.

Finally, you may verify that the recoding was done correctly by computing a correlation coefficient between the initial variable and the new, recoded variable. If the recoding was conducted properly, there should be a perfect negative correlation of -1 between the variables.

Cautions

Recent evidence suggests that the use of scales with a mixed format, traditionally and reverse coded items on the same scale, can adversely affect the psychometric properties of scales. If participants misinterpret a reverse coded item, then the score used in the analysis is not a true measure; in fact, the score is in error by the maximum amount detected on the scale (i.e., a score of 1 instead of a score of 5). The inclusion of reverse coded items on an instrument necessitates the assumption that participants both interpret and respond to the items in the same manner, irrespective of item wording. However, researchers have found that mixed format scales may potentially confound factor structure; result in a separate factor for the

reverse coded items; result in significant differences from responses to traditionally worded items; reduce scale reliability; and result in less accurate responses, therefore hindering the validity of the results.

If participants do interpret and/or react differently to negatively worded items, then what contributes to this difference? Participants may overlook the word or phrase that reversed the meaning of the item (e.g., The instructor did *not* present concepts clearly) or may have difficulty with the mental processing required to strongly disagree to a reverse coded item in an effort to give a positive rating to the construct under measurement (e.g., The instructor is impatient with students). To date, researchers have attributed differential response patterns to mixed item wordings to age, careless responses, culture, educational level, insufficient cognitive ability, instrument wording, interest in topic, personality traits, reading ability, and the actual measurement of a different construct.

Therefore, when using reverse coded items, researchers must determine if the potential benefits outweigh potential costs. Recommendations made from research in the area of scales with a mixed format include using only positively stated items, avoiding the comparison of scores from scales with different numbers of reverse coded items, and conducting separate analyses for the reverse coded items.

—Gail Weems

Further Reading

- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively-worded stems. *Educational and Psychological Measurement*, 60(3), 361–370.
- Benson, J., & Wilcox, S. (1981, April). *The effect of positive and negative item phrasing on the measurement of attitudes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles. (ERIC Document Reproduction Service No. ED204404)
- Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports*, 88, 497–500.

- Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly*, 14(3), 31–36.
- Schmitt, N., & Stults, D. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 4, 367–373.
- Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, 51(1), 67–78.
- Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, 34(3), 166–176.
- Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. T. (in press). The role of reading comprehension in responses to positively- and negatively-worded items on rating scales. *Evaluation & Research in Education*.
- Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. C. (2003). Profiles of respondents who respond inconsistently to positively- and negatively-worded items on rating scales. *Evaluation and Research in Education*, 17(1), 45–60.
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents with different response patterns to positively- and negatively-worded items on rating scales. *Assessment and Evaluation in Higher Education*, 28(6), 587–607.

REYNOLDS, CECIL R. (1952–)

Cecil R. Reynolds earned his doctoral degree from the University of Georgia in 1978, with a major in school psychology and minors in statistics and clinical neuropsychology. Dr. Reynolds studied under Alan Kaufman, author of various tests, including the Kaufman Assessment Battery for Children (KABC).

Dr. Reynolds was a faculty member at the University of Nebraska–Lincoln (1978–1981), where he served as Associate Director and Acting Director of the Buros Institute of Mental Measurement following the death of its founder, Oscar Buros. Dr. Reynolds is currently a professor in the School Psychology Program at Texas A&M University in College Station, Texas, where he has been a professor since 1981.

Dr. Reynolds has authored more than 300 scholarly publications and has been the author or editor of

39 books, including *The Clinician's Guide to the BASC*, *Clinical Applications of Continuous Performance Tests*, *Handbook of School Psychology*, the *Encyclopedia of Special Education*, and the *Handbook of Clinical Child Neuropsychology*. Dr. Reynolds has been actively involved in test development since 1978, working as research or psychometric consultant for tests such as the KABC, Kaufman Adolescent and Adult Intelligence Test for Adolescents and Adults, and the revision of the Detroit Tests of Learning Aptitude. Dr. Reynolds is also the author of several widely used tests of personality and behavior, including the Behavior Assessment System for Children and the Revised Children's Manifest Anxiety Scale. He is also senior author of the Test of Memory and Learning, the Clinical Assessment Scales for the Elderly, the forthcoming Elderly Memory Schedule, and co-author of several computerized test interpretation systems. He is senior author of the Reynolds Intellectual Assessment Scales.

Dr. Reynolds has been president of various professional organizations, including the American Board of Professional Neuropsychology. He is also a past president of the National Academy of Neuropsychology; APA Division 5 (Evaluation, Measurement, and Statistics); and APA Division 40 (Clinical Neuropsychology). Most recently, in 2004 and 2005, he was President of APA Division 16 (School Psychology). Dr. Reynolds has received numerous honors and awards. Among his awards are the Distinguished Research Scholar Award from Texas A&M University (1995), Senior Scientist Award by the American Psychological Association, Division 16 (1999), Distinguished Clinical Neuropsychologist Award by the National Academy of Neuropsychology (2000), Lifetime Achievement Award in Neuropsychology by the National Association of School Psychologists Neuropsychology Interest Group (2003), and Distinguished Alumnus Award for Lifetime Achievement by the University of Georgia (2005).

—Wilda Laija-Rodriguez

Further Reading

Cecil Reynolds and BASC-2: <http://www.agsnet.com/psych/oct04a.asp>

ROBERTS APPERCEPTION TEST FOR CHILDREN

The Roberts Apperception Test for Children (RATC), published by Western Psychological Services, is a projective test used to assess children's psychological development. Its primary purpose is to assess children's perceptions of common interpersonal situations as an aid to general personality description and clinical decision making. Interpretation of the RATC is based on the "projective hypothesis," an assumption that children, when presented with ambiguous drawings of children and adults in everyday interaction, will project their characteristic thoughts, concerns, conflicts, and coping styles into the stories they create. During the test, the individual is presented with a series of drawings and is asked to create stories describing what is happening in each situation, what led up to it, and how it will end. Usually, the child is asked to explain what the main characters are thinking and feeling. For example, a child is pictured sitting at a desk, surrounded by books and papers, apparently engaged in homework. In another picture, a child is pictured kneeling in front of, and clasping, a female figure who has her arms around the child in an apparently comforting posture.

The test's standardized and formally coded content scales indicate where the child is on a continuum of social understanding. Typically, as children become more socially experienced, their stories reflect greater awareness of social convention, more differentiated themes, and clearer resolution of themes and conflicts. Moreover, the inclusion of clinical scales calls attention to the likely presence of social and emotional problems that are outside the norm. Thus, the RATC assesses two independent dimensions: adaptive social perception and the presence of maladaptive or atypical social perception. The RATC takes approximately 20–30 minutes to administer when all stimulus cards are presented. After recording the stories, the clinician scores the responses for the presence or absence of specific characteristics. Stories are rated on seven scales: Theme Overview, Problem Identification, Emotion, Available Resources, Resolution, Outcome, and Unusual or Atypical Responses. Scores are

plotted onto a profile sheet that shows *T*-score equivalents of raw scores and a shaded area between 40*T* and 60*T* denoting the "normal range."

The RATC was standardized on a sample of 200 children who had been described as "generally well adjusted" by their teachers. This standardization sample has been criticized by a number of researchers who have shown that the RATC is of questionable value for distinguishing between clinical and nonclinical populations, and they have suggested that the norms not be used for clinical diagnosis. There have also been debates about the reliability and validity research on the RATC, despite it being the second most popular projective test used with children in the USA. However, in the Roberts-2 (the second edition of the RATC, which was published in late 2005), new norms, grouped by age and sex, are based on a sample of 1,000 children and adolescents (age extended to 18 years), and this test is better representative in terms of gender, ethnicity, and parental education than the original sample. The Roberts-2 also includes data on a clinically referred sample of more than 500 children and adolescents. The authors suggest that this new evidence supports the use of the Roberts-2 in both research on the development of social understanding in normal children, and the clinical assessment of children experiencing adjustment problems. The Roberts-2 also includes new test pictures that feature current hair and clothing styles but retain the thematic content of the original pictures, as well as three parallel versions of the test pictures—one showing White children, one featuring Black children, and a third depicting Hispanic children.

—Fran Vertue

Further Reading

- Bell, N. L., & Nagle, R. J. (1999). Interpretive issues with the Roberts Apperception Test for Children: Limitations of the standardization group. *Psychology in the Schools*, 36, 277–283.
- Finch, A. J., & Belter, R. W. (1993). Projective techniques. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child and adolescent assessment* (pp. 224–238). Boston: Allyn & Bacon.
- Louw, A. E., & Ramkissoon, S. (2002). The suitability of the Roberts Apperception Test for Children (RATC), The

House-Tree-Person (H-T-P), and Draw-a-Person (D-A-P) scales in the identification of child sexual abuse in the Indian community: An exploratory study. *Southern African Journal of Child and Adolescent Mental Health*, 14, 91–106.

McArthur, D. S., & Roberts, G. E. (1982). *Roberts Apperception Test for Children (RATC) manual*. Los Angeles: Western Psychological Services.

Roberts Apperception Test for Children product information: www.wpspublish.com

RORSCHACH INKBLOT TEST

The Rorschach is the well-known inkblot technique, named for its inventor, a Swiss psychiatrist who noticed that mental patients responded to inkblots differently from normal people. Published by Hans Huber of Bern, the 10 inkblots are among the most widely used and most roundly criticized assessment techniques in psychology. Subjects are asked to find images in the ink, and their response tendencies are inferred from what they saw and from how they used the ink.

The Rorschach has three basic uses. One, the inkblots are standardized stimuli that invite a wide variety of responses. Complete and leisurely inspection of the stimulus that occasions a response is a rarity for psychologists and allows for careful consideration of the functional relationship between the two. The psychologist infers what kinds of variables were operating on the subject to produce this particular response to this particular stimulus. Research has shown that managing the gradient from abstraction (they do represent some things) to concreteness (they are only inkblots) may be particularly difficult for seriously disturbed people. Detractors note that research has also shown that interpretations sometimes say more about the psychologist than the subject.

Two, comparison of a subject's responses with what the inkblots actually look like provides a test of the subject's ability to perceive reality accurately. Detractors note that it is hard to say what something actually looks like. Any test of perceptual accuracy, however, will either produce little variation among subjects or raise questions about the psychologist's subjective evaluation. Clinicians frequently make

judgments about people based on subjective evaluations; the standardized stimuli of the Rorschach at least allow for a second opinion.

Three, the Rorschach is widely used as a nomothetic device, meaning that responses are coded for how the ink was used, the content that was seen, and numerous other variables. An individual's codes are then compared with population averages, and inferences are drawn about subjects by comparing their performance with most people's. Nomothetic use has proliferated since John E. Exner published the Comprehensive System (CS) in 1974, with its extensive research base. Critics have attacked the CS as not meeting nomothetic standards with respect to reliability, norming, and validity.

Few critics take issue with the utility of examining a subject's responses to ambiguous stimuli under controlled conditions. Many Rorschach interpretations, however, go far beyond this. The CS, for example, derives 113 codes and 32 indexes, ratios, and percentages from as few as 14 answers. The Rorschach takes a relatively long time to learn, and critics are also concerned that this investment makes practitioner constraint more difficult.

—Michael Karson

Further Reading

- Exner, J. (2003). *The Rorschach: A comprehensive system* (4th ed.). New York: Wiley.
- Karson, M., & Kline, C. (2004, April 4). Two interpretations of Jim Wood's specimen Rorschach protocol. *WebPsych Empiricist*. Retrieved from <http://home.earthlink.net/~rkmck/vault/karson04/karson04.pdf>
- Wood, J., Nezworski, M. T., Lilienfeld, S., & Garb, H. (2003). *What's wrong with the Rorschach? Science confronts the controversial inkblot test*. San Francisco: Jossey-Bass.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Garb, H. N., Wood, J. M., & Nezworski, M. T. (2000). Projective techniques and the detection of child sexual abuse. *Child Maltreatment*, 5(2), 161–168. Projective techniques such as the **Rorschach inkblot test** are sometimes used to detect child sexual abuse. A previously conducted meta-analysis on this topic excluded nonsignificant results. In

this article, a reanalysis of that data is presented. Dr. Garb and his colleagues conclude that projective techniques should not be used to detect child sexual abuse. Many of the studies purportedly demonstrating validity are flawed, and none of the projective test scores have been well replicated.

R_V AND CONGRUENCE COEFFICIENTS

The congruence coefficient was first introduced by Burt under the name of *unadjusted correlation* as a measure of the similarity of two factorial configurations. The name *congruence* coefficient was later tailored by Tucker. The congruence coefficient is also sometimes called a *monotonicity* coefficient. The congruence coefficient takes values between -1 and $+1$.

The R_V coefficient was introduced by Escoufier as a measure of similarity between squared symmetric matrices (specifically, positive semi-definite matrices; see the Appendix section at the end of this entry for a proof) and as a theoretical tool to analyze multivariate techniques. The R_V coefficient is used in several statistical techniques such as STATIS and DISTATIS. In order to compare rectangular matrices using the R_V coefficient, the first step is to transform them into square matrices. The R_V coefficient takes values between 0 and $+1$ (because it is used with positive semi-definite matrices).

These coefficients are similar to the correlation coefficient and are sometimes called *vector* or *matrix* correlation coefficients. This is a potentially misleading appellation because these coefficients are *not* correlation coefficients because contrary to the correlation coefficient, the mean of the observations is not subtracted prior to the computation.

The computational formulas of these coefficients are identical, but their usage and theoretical foundations differ. Also, their sampling distributions differ because of the types of matrices with which they are used.

Notations and Computational Formulas

Let $\mathbf{X}$ be an $I \times J$ matrix and $\mathbf{Y}$ be an $I \times K$ matrix. The *vec* operation transforms a matrix into a vector whose

entries are the elements of the matrix. The *trace* operation applies to square matrices and gives the sum of the diagonal elements.

The congruence coefficient is defined when both matrices have the same number of rows and columns (i.e., $J = K$). These matrices can represent factor loadings (i.e., factors by items) or factor projections (i.e., observations by factors). The congruence coefficient is denoted φ or sometimes r_c , and it can be computed with three different equivalent formulas:

$$\varphi = r_c = \frac{\sum_{i,j} x_{i,j} y_{i,j}}{\sqrt{\left(\sum_{i,j} x_{i,j}^2\right) \left(\sum_{i,j} y_{i,j}^2\right)}} \quad (1)$$

$$= \frac{\text{vec}\{\mathbf{X}\}^T \text{vec}\{\mathbf{Y}\}}{\sqrt{(\text{vec}\{\mathbf{X}\}^T \text{vec}\{\mathbf{X}\}) (\text{vec}\{\mathbf{Y}\}^T \text{vec}\{\mathbf{Y}\})}} \quad (2)$$

$$= \frac{\text{trace}\{\mathbf{XY}^T\}}{\sqrt{(\text{trace}\{\mathbf{XX}^T\}) (\text{trace}\{\mathbf{YY}^T\})}}. \quad (3)$$

The R_V coefficient was defined by Escoufier as a similarity coefficient between positive semi-definite matrices. Escoufier and Robert and Escoufier pointed out that the R_V coefficient had important mathematical properties because it can be shown that most multivariate analysis techniques amount to maximizing this coefficient with suitable constraints. Recall, at this point, that a matrix $\mathbf{S}$ is called *positive semi-definite* when it can be obtained as the product of a matrix by its transpose. Formally, we say that $\mathbf{S}$ is positive semi-definite when there exists a matrix $\mathbf{X}$ such that

$$\mathbf{S} = \mathbf{XX}^T. \quad (4)$$

Note that as a consequence of the definition, positive semi-definite matrices are square and symmetric, and their diagonal elements are always larger or equal to zero.

If we denote by $\mathbf{S}$ and $\mathbf{T}$ two positive semi-definite matrices of same dimensions, the R_V coefficient between them is defined as

$$R_V = \frac{\text{trace}\{\mathbf{S}^T \mathbf{T}\}}{\sqrt{(\text{trace}\{\mathbf{S}^T \mathbf{S}\}) \times (\text{trace}\{\mathbf{T}^T \mathbf{T}\})}}. \quad (5)$$

This formula is computationally equivalent to

$$R_V = \frac{\text{vec}\{\mathbf{S}\}^T \text{vec}\{\mathbf{T}\}}{\sqrt{(\text{vec}\{\mathbf{S}\}^T \text{vec}\{\mathbf{S}\})(\text{vec}\{\mathbf{T}\}^T \text{vec}\{\mathbf{T}\})}}. \quad (6)$$

$$= \frac{\sum_i^I \sum_j^I s_{i,j} t_{i,j}}{\sqrt{\left(\sum_i^I \sum_j^I s_{i,j}^2\right) \left(\sum_i^I \sum_j^I t_{i,j}^2\right)}}. \quad (7)$$

For rectangular matrices, the first step is to transform the matrices into positive semi-definite matrices by multiplying each matrix by its transpose. So, in order to compute the value of the R_V coefficient between the $I \times J$ matrix $\mathbf{X}$ and the $I \times K$ matrix $\mathbf{Y}$, the first step is to compute

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T \text{ and } \mathbf{T} = \mathbf{Y}\mathbf{Y}^T. \quad (8)$$

If we combine Equations 5 and 8, we find that the R_V coefficient between these two rectangular matrices is equal to

$$R_V = \frac{\text{trace}\{\mathbf{X}\mathbf{X}^T \mathbf{Y}\mathbf{Y}^T\}}{\sqrt{(\text{trace}\{\mathbf{X}\mathbf{X}^T \mathbf{X}\mathbf{X}^T\}) \times (\text{trace}\{\mathbf{Y}\mathbf{Y}^T \mathbf{Y}\mathbf{Y}^T\})}} \quad (9)$$

The comparison of Equations 3 and 9 shows that the congruence and the R_V coefficients are equivalent only in the case of positive semi-definite matrices.

From a linear algebra point of view, the numerator of the R_V coefficient corresponds to a scalar product

between positive semi-definite matrices, and therefore gives to this set of matrices the structure of a vector space. Within this framework, the denominator of the R_V coefficient is called the Frobenius, or Schur, or Hilbert-Schmidt matrix scalar product, and the R_V coefficient is a *cosine* between matrices. This vector space structure is responsible for the mathematical properties of the R_V coefficient.

Sampling Distributions

The congruence and the R_V coefficients quantify the similarity between two matrices. An obvious practical problem is to be able to perform statistical testing on the value of a given coefficient. In particular, it is often important to be able to decide if a value of a coefficient could have been obtained by chance alone. To perform such statistical tests, we need to derive the sampling distribution of the coefficient under the null hypothesis (i.e., in order to test if the population coefficient is null). More sophisticated testing requires deriving the sampling distribution for different values of the population parameters. So far, analytical methods have failed to characterize such distributions, but computational approaches have been used with some success. Because the congruence and the R_V coefficients are used with different types of matrices, their sampling distributions are likely to differ, and so, work done with each type of coefficient has been carried independently of the other.

Congruence Coefficient

Recognizing that analytical methods were unsuccessful, Korth and Tucker decided to use Monte Carlo simulations to gain some insights into the sampling distribution of the congruence coefficient. Their work was completed by Broadbooks and Elmore. From this work, it seems that the sampling distribution of the congruence coefficient depends upon several parameters, including the original factorial structure and the intensity of the population coefficient, and therefore, no simple picture emerges, but some approximations can be used. In particular, for testing that a congruence coefficient is null in the

population, an approximate conservative test is to use Fisher Z transform and to treat the congruence coefficient like a coefficient of correlation. Broadbooks and Elmore provide tables for population values different from zero. With the availability of fast computers, these tables easily can be extended to accommodate specific cases.

R_V Coefficient

Statistical approaches for the R_V coefficient have focused on permutation tests. In this framework, the permutations are performed on the entries of each column of the rectangular matrices $\mathbf{X}$ and $\mathbf{Y}$ used to create the matrices $\mathbf{S}$ and $\mathbf{T}$. Interestingly, work by Kazi-Aoual, Hitier, Sabatier and Lebreton has shown that the mean and the variance of the permutation test distribution can be computed directly from $\mathbf{S}$ and $\mathbf{T}$.

The first step is to derive an index of the dimensionality or rank of the matrices. This index, denoted β_S (for matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$), is also known as ν in the brain imaging literature, where it is called a *sphericity* index and is used as an estimation of the number of degrees of freedom for multivariate tests of the general linear model. This index depends upon the eigenvalues of the $\mathbf{S}$ matrix, denoted $s\lambda_\ell$, and it is defined as

$$\beta_S = \frac{\left(\sum_{\ell}^L s\lambda_{\ell}\right)^2}{\sum_{\ell}^L s\lambda_{\ell}^2} = \frac{\text{trace}\{\mathbf{S}\}^2}{\text{trace}\{\mathbf{S}\mathbf{S}\}}. \quad (10)$$

The mean of the set of permuted coefficients between matrices $\mathbf{S}$ and $\mathbf{T}$ is then equal to

$$E(R_V) = \frac{\sqrt{\beta_S \beta_T}}{I - 1}. \quad (11)$$

The case of the variance is more complex and involves computing three preliminary quantities for each matrix. The first quantity is denoted δ_S (for matrix $\mathbf{S}$), and it is equal to

$$\delta_S = \frac{\sum_i^I s_{i,i}^2}{\sum_{\ell}^L s\lambda_{\ell}^2}. \quad (12)$$

The second one is denoted α_S , for matrix $\mathbf{S}$, and is defined as

$$\alpha_S = I - 1 - \beta_S. \quad (13)$$

The third one is denoted C_S (for matrix $\mathbf{S}$) and is defined as

$$C_S = \frac{(I - 1)[I(I + 1)\delta_S - (I - 1)(\beta_S + 2)]}{\alpha_S(I - 3)}. \quad (14)$$

With these notations, the variance of the permuted coefficients is obtained as

$$V(R_V) = \alpha_S \alpha_T \times \frac{2I(I - 1) + (I - 3)C_S C_T}{I(I + 1)(I - 2)(I - 1)^3}. \quad (15)$$

The sampling distribution of the permuted coefficients is relatively similar to a normal distribution (even though it is, in general, *not* normal), and therefore, we can use a Z criterion to perform null hypothesis testing or to compute confidence intervals. For example, the criterion

$$Z_{R_V} = \frac{R_V - E(R_V)}{\sqrt{V(R_V)}} \quad (16)$$

can be used to test the null hypothesis that the observed value of R_V was due to chance.

An Example

As an example, we will use two scalar product matrices from the STATIS example entry (Experts 1 and 3). These matrices are listed as follows:

$$\mathbf{S} = \begin{bmatrix} 29.56 & -8.78 & -20.78 & -20.11 & 12.89 & 7.22 \\ -8.78 & 2.89 & 5.89 & 5.56 & -3.44 & -2.11 \\ -20.78 & 5.89 & 14.89 & 14.56 & -9.44 & -5.11 \\ -20.11 & 5.56 & 14.56 & 16.22 & -10.78 & -5.44 \\ 12.89 & -3.44 & -9.44 & -10.78 & 7.22 & 3.56 \\ 7.22 & -2.11 & -5.11 & -5.44 & 3.56 & 1.89 \end{bmatrix} \quad (17)$$

and

$$\mathbf{T} = \begin{bmatrix} 11.81 & -3.69 & -15.19 & -9.69 & 8.97 & 7.81 \\ -3.69 & 1.81 & 7.31 & 1.81 & -3.53 & -3.69 \\ -15.19 & 7.31 & 34.81 & 9.31 & -16.03 & -20.19 \\ -9.69 & 1.81 & 9.31 & 10.81 & -6.53 & -5.69 \\ 8.97 & -3.53 & -16.03 & -6.53 & 8.14 & 8.97 \\ 7.81 & -3.69 & -20.19 & -5.69 & 8.97 & 12.81 \end{bmatrix}. \quad (18)$$

We find the following value for the R_V coefficient:

$$R_V = \frac{\sum_i^I \sum_j^I s_{i,j} t_{i,j}}{\sqrt{\left(\sum_i^I \sum_j^I s_{i,j}^2 \right) \left(\sum_i^I \sum_j^I t_{i,j}^2 \right)}} = \frac{(29.56 \times 11.81) + (-8.78 \times -3.69) + \dots + (1.89 \times 12.81)}{\sqrt{[(29.56)^2 + (-8.78)^2 + \dots + (1.89)^2][(11.81)^2 + (-3.69)^2 + \dots + (12.81)^2]}} = .79. \quad (19)$$

To test the significance of a value of $R_V = .79$, we first compute the following quantities:

$$\begin{aligned} \beta_S &= 1.0954 & \alpha_S &= 3.9046 & \delta_S &= 0.2951 & C_S &= 1.3162 \\ \beta_T &= 1.3851 & \alpha_T &= 3.6149 & \delta_T &= 0.3666 & C_T &= 0.7045 \end{aligned} \quad (20)$$

Plugging these values into Equations 11, 15, and 16, we find

$$E(R_V) = 0.2464, V(R_V) = 0.0422, \text{ and } Z_{R_V} = 2.66. \quad (21)$$

Assuming a normal distribution for the Z_{R_V} gives a p value of .0077, which would allow for the rejection

of the null hypothesis for the observed value of the R_V coefficient.

Appendix

The R_V coefficient takes values between 0 and 1 for positive semi-definite matrices.

Let $\mathbf{S}$ and $\mathbf{T}$ be two positive semi-definite matrices. From the Cauchy-Schwartz inequality, we know that the absolute value of the numerator is always smaller or equal to the denominator (and so R_V is smaller than 1); therefore we only need to prove that the numerator of the R_V coefficient is positive or null. This amounts to showing that

$$\text{trace}\{\mathbf{ST}\} \geq 0. \quad (22)$$

Because $\mathbf{T}$ is positive semi-definite, the left part of Equation 22 can be rewritten as

$$\text{trace}\{\mathbf{T}^{\frac{1}{2}}\mathbf{ST}^{\frac{1}{2}}\}. \quad (23)$$

Because $\mathbf{S}$ is positive semi-definite, the matrix $(\mathbf{T}^{\frac{1}{2}}\mathbf{ST}^{\frac{1}{2}})$ is also positive semi-definite, and therefore all its eigenvalues are positive or null. Thus, its trace being equal to the sum of its eigenvalues is also positive or null, and this completes the proof.

—Hervé Abdi

See also DISTATIS; Eigendecomposition; Eigenvalues; Multiple Factor Analysis; STATIS

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Bedeian, A. G., Armenakis, A. A., & Randolph, W. A. (1988). The significance of congruence coefficients: A comment and statistical test. *Journal of Management*, 14, 559–566.
- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer Verlag.
- Broadbooks, W. J., & Elmore, P. B. (1987). A Monte Carlo study of the sampling distribution of the congruence coefficient. *Educational and Psychological Measurement*, 47, 1–11.
- Burt, C. (1948). Factor analysis and canonical correlations. *British Journal of Psychology, Statistical Section*, 1, 95–106.
- Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29, 751–760.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed. rev.). Chicago: Chicago University Press.
- Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge, UK: Cambridge University Press.
- Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20, 643–656.
- Korth, B. A., & Tucker, L. R. (1976). The distribution of chance congruence coefficients from simulated data. *Psychometrika*, 40, 361–372.
- Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64.
- Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics*, 25, 257–265.
- Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T. Næs & E. Risvik (Eds.), *Multivariate analysis of data in sensory sciences*. New York: Elsevier.
- Tucker, L. R. (1951). *A method for the synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.
- Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—Again. *NeuroImage*, 2, 173–181.

S

Lottery is a tax on people who are bad at math.

—Author unknown

SAMPLE

It is very rare that all units in a population can be measured. More commonly, a selection of units from a population is chosen, and only these selected units are measured. A population may be all the people who live in a city, all the children currently enrolled in a school, all the trees of a particular species in a forest, or all the sales transactions of a shop this month. The selection of elements from a population is called a *sample*. In these examples, the samples would be samples of people, of children, of trees, or of transactions.

Samples are selected according to a *sampling scheme*. The size of the sample to be selected should be defined in the scheme. Summary statistics are calculated from the sample and are used to estimate population parameters. Choosing the best sample scheme and selecting the most appropriate method to estimate population parameters are the main interests in research in sample surveys.

The most elementary sample scheme is simple random sampling. With a finite population of N units, the number of ways the population can be arranged into subsets of n distinct units can be calculated as follows:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

The subsets represent all the possible samples of size n distinct units that could be taken from a finite population of size N . In simple random sampling, all these possible samples have an equal chance of being selected. For example, consider a population in which $N = 5$. For convenience, the units are labeled A, B, C, D, and E. Ten possible samples of size 3 could be taken:

A	B	C
A	B	D
A	B	E
B	C	D
B	C	E
C	D	E
A	C	D
A	C	E
A	D	E
B	D	E

Samples can be chosen so that each unit appears in the sample only once, or they can be chosen so that each unit can in the sample more than once. When units can appear only once in the sample (as in the example above), the process is called *sampling without replacement*, and when units can appear more than once, the process is *sampling with replacement*. Drawing names or numbers out of a hat is an example

of sampling without replacement, provided none of the names or numbers is returned to the hat. Selecting units from the population by means of a random table of numbers that correspond to the unit labels is an example of sampling with replacement because the same random number can appear more than once, and hence the same population unit can be selected more than once. Sampling without replacement may be more informative than sampling with replacement because every unit in the sample appears only once and brings new information. If the same unit appeared twice, the second occurrence of the unit would not bring new information.

—Jennifer Ann Brown

Further Reading

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Thompson, S. K. (1992). *Sampling*. New York: Wiley.

SAMPLE SIZE

Determination of sample size involves methods for deciding how much data should be collected in a statistical study. The sample size is chosen to meet one or more goals, which could relate to precision of estimation, power, cost, or some other criterion. Sample size is not equally important in all studies. It is relatively unimportant in planning a pilot study or in cases in which additional data, if needed, can be collected quickly with minimal planning. However, when one is planning a definitive study, sample size determination can be critical, especially when the process of acquiring subjects or participants, materials, measurements, budgetary support, or grant funding is lengthy.

It is important to know that in sample-size planning, science must come before statistics. That is, we first must establish the scientific goals of the study and then address the statistical issues in meeting those goals. There are often a number of intermediate steps—for example, deciding the target population, the way measurements are to be made, and so

forth. Typically, the scientific goals are stated in terms of a difference (or other measure of effect size) expressed in the same units as the planned measurements; we choose an effect that would be deemed important based on substantive considerations. Once the scientific goals are established, we can address the statistical design to be used and, finally, the sample size.

It is seldom possible to do a definitive statistical study in one step. Unless there is substantial past experience with the measurements to be used and the participant populations to be studied, a pilot study is necessary. The results of the pilot study provide estimates of variances and other quantities needed as inputs to the calculations to determine sample size. A pilot study also helps ensure that the planned scheduling and procedures will actually work and that the training program is adequate for those conducting interviews and collecting data.

Sample Size for Specified Confidence Interval Width

Perhaps the simplest approach to sample-size estimation is to specify the desired width of the confidence interval for some parameter of interest. For example, suppose that an achievement test is to be administered to selected ninth graders in an alternative instructional program, and we want to estimate the mean achievement, μ , to within a margin of error of ± 5 points (with 95% confidence); past experience with this achievement test, when administered to other populations, suggests that the standard deviation (SD) of the scores is about 15. We plan to use the usual confidence interval based on the t distribution. For a confidence coefficient of $1 - \alpha$, the sample-size formula is

$$n = \left(\frac{t_{\alpha/2, n-1} \sigma}{\delta} \right)^2,$$

where

$t_{\alpha/2, n-1}$ is the critical value for an upper tail area of $\alpha/2$ on the t distribution with $n - 1$ degrees of freedom (df),

σ is the population *SD*, and

δ is the desired margin of error.

The t critical value depends on n , but not strongly; we can start with, say, 30 *df*, and iterate once or twice until it stabilizes. In the example, we use $\alpha = .05$, start with $t_{\alpha/2, n-1} = t_{.025, 30} = 2.042$ (from standard t tables), $\sigma = 15$, and $\delta = 5$, to obtain a starting value of

$$n = \left(\frac{2.042 \times 15}{5} \right)^2 = 37.53,$$

which would round to $n = 38$. For this n , we should use 37 *df* or $t_{.025, 37} = 2.026$, so in a new iteration, we get $n = (2.026 \times 15/5)^2 = 36.94$, rounded to $n = 37$. Correcting to 36 *df* will not change the result enough to matter, and our conclusion is that 37 participants are enough to estimate the mean to within 5 points. Strictly speaking, fractional results for n should always be taken to the next higher whole number; but in practice, it makes little difference except when n is very small.

Sample Size to Meet Power Requirements

In many studies, the primary statistical inferences will be the results of one or more significance tests. In those cases, it is common practice to choose sample size so that the power of the test has at least a specified value, given a specified significance level and effect size of scientific or clinical interest. Typically, for significance level $\alpha = .05$, one chooses a target power of $\pi = .8$ or $\pi = .9$. One rationale for such a choice is to weigh the severity of a Type I error (rejecting the null hypothesis when it is true) and a Type II error (failing to detect an effect of the specified size). For example, if one chooses $\alpha = .05$, and $\pi = .8$, the latter is equivalent to a Type II error probability of $\beta = 1 - \pi = .20$, so there is an implicit statement that a Type I error is 4 times worse than a Type II error. Another way to look at the same situation would be to specify $\alpha = .05$ and $\pi = .8$, then specify the effect size for which a Type II error is one fourth

as bad as a Type I error; or to choose $\pi = .95$ and choose a (larger) effect size such that failing to detect it would be equally bad as a Type I error.

First, consider a two-sided hypothesis test of the form

$$H_0: \theta = \theta_0 \text{ versus } H_1: \theta \neq \theta_0,$$

where θ is a parameter of interest and θ_0 is a null value, and a test statistic of the form

$$z = \frac{\hat{\theta} - \theta_0}{c\sigma/\sqrt{n}} \text{ or } t = \frac{\hat{\theta} - \theta_0}{cs/\sqrt{n}},$$

where

$\hat{\theta}$ is an estimate of θ ,

$\sigma(c/\sqrt{n})$ is the standard error of $\hat{\theta}$ (here, c is some known value), and

s is an estimate of σ .

Then an approximate sample-size formula for achieving power π when $\theta = \theta_1$ at significance level α is

$$n = c^2 \left(\frac{(z_{\alpha/2} + z_{1-\pi})\sigma}{\theta_1 - \theta_0} \right)^2.$$

This formula is exact for the z test, but it underestimates sample size somewhat for the t test. In the latter case, a better approximation is obtained using t critical values in place of the z critical values, but the method is still only approximate.

For illustration, suppose that we plan to do a t test of independent samples to compare the mean reaction times of laboratory rats under two types of conditioning. Based on pilot data collected using the same instrumentation, we estimate that the *SD* of reaction times is 320 msec. It is decided that a difference of means of 250 msec or more would be regarded as clinically important, and we want to be able to detect such a difference with a power of .90 based on a test with significance level .02. The corresponding hypotheses are

$$H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 \neq \mu_2,$$

and the t statistic, based on independent samples of equal size n and sample means $\bar{y}_1$ and $\bar{y}_2$ and SDs s_1 and s_2 , is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{2/n}},$$

where $s = \sqrt{(s_1^2 + s_2^2)/2}$ is the pooled SD. By examining the t statistic, we see that $c = \sqrt{2}$, and hence the required sample size is

$$\begin{aligned} n &= \left(\sqrt{2}\right)^2 \left(\frac{(z_{.01} + z_{.10})\sigma}{\mu_1 - \mu_2}\right)^2 \\ &= 2 \left(\frac{(2.326 + 1.282) \times 320}{250}\right)^2 \\ &= 42.65; \end{aligned}$$

that is, we need 43 rats in each group to meet the stated goals. Moreover, since we know that the formula underestimates sample size, it makes sense to adjust it upward slightly, say to 44 or 45.

Use of Software

Most statistical packages provide some capabilities for sample-size calculations, and some stand-alone sample-size programs are available commercially or for free over the Internet. Some of this software uses approximations such as the formulas above for t tests, others use exact calculations, and some use simulation methods, especially in especially complex situations. Figure 1 shows how the previous example can be done using Piface software, available online. Piface does exact calculations, and it is found that with $n = 44$, a power of nearly .9 is achieved.

Besides saving labor and producing more-reliable results for more situations, sample-size software often affords some flexibility in the way a problem is defined. For example, suppose that in the previous example, the sample size of 44 per treatment is not tractable in terms of available time and resources but that $n = 30$ is considered reasonable. By manipulating the options in the Piface dialog, we can easily find what power is achieved for the same effect size (the

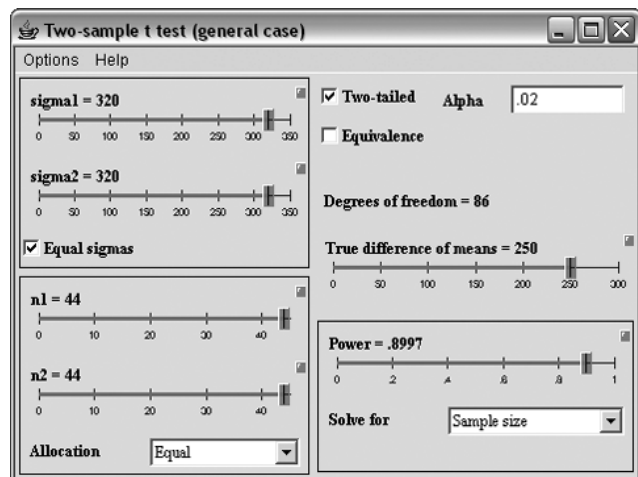


Figure 1 Piface Dialog for a Two-Sample t Test (Reaction-Time Example)

answer is about .735, which is still reasonable). Or we could ask the software to solve for effect size instead of sample size, as shown in Figure 2; it turns out that we would be able to detect a difference of means of about 305 msec with a power of .9.

Analysis of Variance

In the analysis of variance (ANOVA), an F test is constructed for a null hypothesis that k means are equal:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k.$$

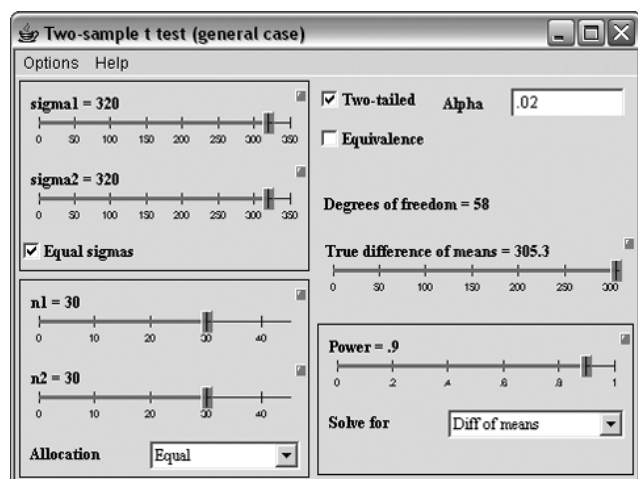


Figure 2 Using the Piface Dialog to Solve for Effect Size

For studying the power of the F test, we must specify an alternative hypothesis H_1 of substantive interest, that is, a set of values of μ_j . The power of the test depends on the SD of these μ_j . For example, when $k = 8$, consider the following two sets of μ_j : {5, 25, 25, 25, 25, 25, 25, 45} and {15, 15, 15, 15, 35, 35, 35, 35}. Since these have the same SD , the power of the ANOVA F test is the same when either of these sets is specified as the alternative hypothesis of interest, even though the smallest and largest means differ by 40 in the first set and only 20 in the second. Unless a particular relationship among the means is contemplated in advance as a likely outcome, it can be quite difficult to discuss a relevant effect size for the F test at any kind of practical level. The best strategy is probably to construct examples of patterns of means that would be detectable.

An alternative approach, which may be more accessible for discussion purposes, would be to work in terms of effect sizes for pairwise comparisons or contrasts of means. We would then arrive at a suitable sample size for achieving a specified power for a particular comparison of two means. Typically, some type of correction is made for multiple testing, such as a Bonferroni correction or Tukey's honestly significant difference. A similar approach extends to multi-factor ANOVA. For illustration, Figure 3 shows the Piface dialog for a two-factor experiment with factors "treatment" and "dose," where it is desired to be able to detect a difference of 6.0 or more between two cell means. Tukey's honestly significant difference is used to compare the five treatment means at the same dose, and we find that when the error SD is $\sigma = 3.89$, a sample size of 11 observations per cell is adequate to exceed a power of .80.

Sample-Size Allocation

Another aspect of sample size is that of allocation; that is, when there are several groups and we wish to collect N observations total, how should these N

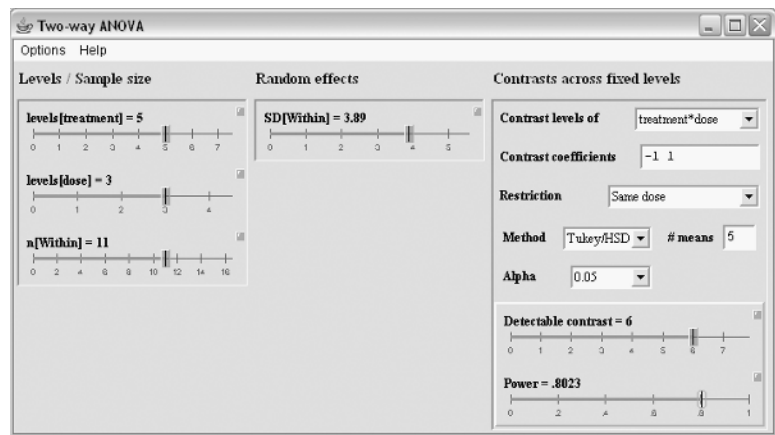


Figure 3

Piface Dialog for Comparing Cell Means in a Two-Factor ANOVA

observations be distributed among the groups? In a comparison of two independent sample means, each group should be allocated $N/2$ observations if an ordinary pooled t test is used; however, if we cannot assume that the SD s are equal, the test should be done using the Satterthwaite approximation, and the sample sizes should be allocated in proportion to the SD s; for example, if one SD is twice as large as the other, the group with the larger SD should get twice as many observations.

Effect-Size Conventions

The approach emphasized here requires one to specify an effect size in terms of the actual units of measurement; accordingly, one also needs an idea of the error SD , perhaps based on pilot data. Another popular approach is to specify effect size in SD units. For example, in a two-sample t test, a difference of half an SD is regarded as a "medium" effect size. This convention is based on a survey of the social science literature conducted by Cohen. The advantage of this approach is that it is easy, requires less thinking, and does not require pilot data. The disadvantage is that it is an elaborate way to choose a predetermined sample size. For example, the sample size for a medium effect is 64 per treatment, regardless of the importance of the research, the reliability of the instrumentation, or the breadth of the population under study. By choosing a medium effect size, we are simply choosing to do a

medium-size study relative to those published in the social science literature. A definitive study requires more stringent scientific standards than that.

—Russell V. Lenth

See also Effect Size; Type I Error; Type II Error

Further Reading

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician*, 55, 187–193.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Odeh, R. E., & Fox, M. (1991). *Sample size choice: Charts for experiments with linear models* (2nd ed.). New York: Marcel Dekker.
- Parker, R. A., & Berman, N. G. (2003). Sample size: More than calculations. *American Statistician*, 57, 166–170.
- Java applets (“Piface”): www.stat.uiowa.edu/~rlenth/Power/ (Either these may be run in a browser with appropriate Java plug-in, or they may be downloaded to run offline. This page also provides links to a number of other online power calculators.)
- Sample-size analysis UnifyPow SAS module, by Ralph O’Brien: www.bio.ri.ccf.org/Power/

SAMPLING DISTRIBUTION OF A STATISTIC

The term *sampling distribution of a statistic* refers to the theoretical, expected distribution for a statistic that would result from taking an infinite number of repeated random samples of size N from some population of interest and calculating the statistic of interest for each sample. The resulting distribution of values is called the sampling distribution of that statistic. For example, the sampling distribution for a sample mean could be constructed by obtaining a random

sample of individuals from a population of interest, computing the mean of observed values, and then repeating this process. That is, one could theoretically obtain repeated random samples of individuals from the population and for each random sample compute the sample mean of values. If this process were repeated indefinitely, the resulting distribution of sample means would be the sampling distribution of the sample mean. This process could be conducted for any given statistic (e.g., sample mean, variance, or correlation).

In practice, a researcher cannot actually create the sampling distribution of a statistic, because it is a theoretical process that requires the creation of an infinitely large number of random samples from some population of interest. However, the concept of the sampling distribution of a statistic is important because it creates the cornerstone for all of inferential statistics. All inferential statistics have in common the process of hypothesis testing. Generally, the process of hypothesis testing entails the assumption that some null hypothesis is true, and then a determination about the likelihood of observing a sample statistic given the null hypothesis. If the observed sample statistic appears very unlikely given the null hypothesis, the analysis results in a p value less than a predetermined critical value, or “alpha level.” In this situation, the null hypothesis is rejected in favor of the alternative hypothesis. For example, a null hypothesis may state that the treatment and control groups from a study have equal means in the population, which would indicate that the treatment has no effect on the outcome of interest. To determine if an observed mean difference is statistically significant, a standardized mean difference is computed between the treatment and control groups. Under the assumption that the treatment has no effect, a large difference between the two groups would be relatively improbable. The p value from this analysis indicates the probability that the null hypothesis is true, given the observed difference. If that p value is small, it indicates that the null hypothesis is probably not true, and the null hypothesis is therefore rejected by the researcher.

This process is familiar enough to most researchers, but it depends entirely on knowing the sampling distribution for the statistic of interest. Without knowing how sample mean differences were distributed in the population, it would be impossible to determine the likelihood of a given mean difference's occurring under the assumption of the null hypothesis. Knowing how a statistic is distributed in the population allows the statistician to make an inference about how likely or unlikely it is that any given sample statistic would have come from some population of interest. This is possible because when a researcher obtains a sample of data and computes some statistic of interest, this value can be treated as one of the possible infinite number of random values that could have been drawn from the population. Then, if the sampling distribution of the statistic is known, the observed sample statistic can be compared to this distribution to determine the probability of its occurring. The sampling distribution of a statistic can be properly thought of as the probability density function for the statistic.

Any inferential statistical analysis requires knowledge of the sampling distribution for that statistic. Different sample statistics, such as sample means, sample mean differences, sample variances, and sample correlations, all follow different sampling distributions, and those sampling distributions may fluctuate depending on the details of a particular analysis, such as sample size, population variability, and population effect sizes. When the parameters that govern the shape of the sampling distribution for a statistic are known, inferences can then be made on the basis of the observed sample statistics.

—William P. Skorupski

See also Hypothesis and Hypothesis Testing; Inferential Statistics

Further Reading

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Boston: Allyn & Bacon.

SAMPLING ERROR

The process of sampling, while essential to the theory and use of inferential statistics, is not perfect, and there is almost always a difference between the sample statistics and the population parameters that are collected. For example, if a population of 1,000 data points has a mean of 14.5, and a sample consisting of 50 data points from that population has an average of 14.1 on the same measure, then the difference is .4. That difference is the error in sampling. The population value is often seen as the true value, and any deviation from that value is seen as an error.

Sampling errors are better understood when they are considered as standard errors, which are simply a standard deviation of the sampling distribution of the statistic under consideration.

For example, the standard error of the mean (*SEM*) is the standard deviation of the sample mean for *n* observations reflecting the variability in sampling error and is computed as follows:

$$SEM = \frac{\sigma}{\sqrt{n}},$$

where

σ is the standard deviation of the original set of scores, and

n is number of sample observations.

For a sample of 9 scores with a standard deviation estimate of 2, the *SEM* is

$$SEM = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{9}} = \frac{2}{3} = .67.$$

Given what we know about any normal distribution, one would expect that 68% of sample means would fall within ± 1 *SEM* or $\pm .67$ units from the grand, or overall, mean.

As one would expect, sampling error decreases as the sample gets larger and the sample more closely approximates the characteristics of the population.

For an infinitely large sample, there would be no sampling error (indeed, σ would equal 0 because the sample and the population values would be identical).

—Neil J. Salkind

SCALING

Scaling is the activity of attempting to measure or quantify psychological attitudes or attributes. The term can also refer to attempts to measure people on some dimension, but here we are concerned with the first-mentioned meaning. The roots of the scaling enterprise go back notably to the German scientist and mystic Gustav Theodor Fechner, who on October 22, 1850 (so he recounted), had an idea that he thought could connect the psychological and the physical realms. Simply phrased, his idea was this: We all know that physical sensory stimuli give rise to perceptions that seem to be quantitative. For example, if two rocks differ in weight by a sufficient amount, one will feel heavier than the other when they are lifted. There were good physical methods for measuring weight. Was there a way to measure the rocks' felt "heavinesses," and could we connect by some mathematical formula the felt heavinesses of the rocks to their weights? Fechner thought he saw a theoretical argument that would connect the physical and psychological magnitudes and concluded that psychological magnitudes were related to the logarithms of their inducing physical intensities. His theoretical argument hinged on some observations of the "just noticeable difference," roughly the minimum difference in stimuli that could be reliably detected. This just noticeable difference appeared in some settings to be a constant proportion of the stimuli's values. So, for example, two rocks could reliably be distinguished in weight if their weights differed by approximately 3%. Combining this observation with the theoretical notion that all just noticeable differences between stimulus weights gave rise to the same constant heaviness difference, Fechner adduced his logarithmic formula connecting physical to psychological magnitudes. The *dol* scale for pain is an application of this approach.

All the schemes for scaling discussed here provide values that are relative. That is, they permit assessment of how different two rated objects are without providing an absolute location on the scale for any single object.

Discriminability Scaling

One early schematic approach to scaling, following on Fechner's interest in discriminability, stems largely from the work of Louis Leon Thurstone and explicitly utilizes some statistical ideas. It relies on the fact that there are comparisons that people make with less than perfect reliability. For example, when repeatedly presented with two objects that differ only a little in physical weight and asked which of the two is the heavier, a person will not always make the same selection. Thus we can theorize that although each object's physical weight is constant, its psychological heaviness undergoes some variation from encounter to encounter. That variation in heaviness is presumed to be the source of the unreliability in a person's selections. The variation takes place in the psychological dimension of heaviness, and it can be presumed to follow a normal distribution. In the simplest version of his work (Case V of the Law of Comparative Judgment), Thurstone added some assumptions that entailed the plausible idea that if stimulus A is judged heavier than stimulus B 80% of the time and stimulus X is judged heavier than stimulus Y 67% of the time, then the heavinesses of A and B differ by more than do the heavinesses of X and Y. A final assumption was that the variation in heaviness had as its form the normal distribution, allowing the translation of the probability with which one stimulus was judged heavier than another into the difference between their heavinesses measured as *z* scores in a normal distribution. Thus a person's performance in an elaborate experiment involving the comparative judgments of many weights can give rise to a scale of heavinesses in which the difference between two weights' scale values accurately predicts the likelihood that one will be judged heavier than the other. Applications of this approach might include having students who had experience with a group of teachers indicate their

favorite in every possible pair of teachers; this could produce scale values for the teachers' "goodnesses." Notice that these scales say nothing about the "absolute" heaviness or goodness of any particular things; they inform us only about the separations among the things being judged.

Magnitude Scaling

Another approach to scaling involves the procedure called *magnitude estimation*, in which people give numerical ratings to stimuli based on some attribute. For example, people might be asked to heft objects of different weights and to assign numbers to the objects so that the numbers are "proportional" to the perceptual experiences of heaviness. This approach presumes that people can make ratio-like comparisons among their sensations in a consistent and unbiased way. (Often, but not necessarily, this is carried out with some particular weight [the *standard*] being assigned a particular numerical value [the *modulus*] and more recently, verbal labels such as "very strong" being attached to some numbers.) The extensive tradition of work in this vein stems largely from the work of Stanley Smith Stevens (who also created the typology of measurement scales—nominal, ordinal, ratio, and interval; he championed the use of magnitude estimation though he did not introduce it). The average (usually the geometric mean) of the numbers assigned to an object is taken to be a measure of its perceived heaviness. In a wide variety of situations involving stimuli that vary in "intensity" (such as lights, sounds, smells, etc.), a simple form of algebraic equation characterizes the relation between the physical intensities of the stimuli and the numbers people assign to them. The numbers people assign vary as some power of the physical intensities of the stimuli, or equivalently, the numbers are a power function of the physical intensities. So, for example, the loudness ratings of 1,000 Hertz tones are linearly related to approximately the cube roots of the sounds' physical intensities (i.e., to the intensities raised to the $1/3$ power, or to intensities with an exponent of $1/3$). Such relations, called power functions, are found in many studies of this sort (e.g., the brightness of lights,

the saltiness of salt solutions, the painfulness of electric shocks). Power function relationships are linear when graphed in double-logarithmic coordinates (i.e., when both axes of the graph [the stimulus intensities and the numbers people say] are logarithmic; this arrangement makes it easy to determine whether a given set of data is consistent with that algebraic form). The power to which the intensities are raised is equal to the slope of the line that connects the data in double-logarithmic coordinates and is usually the main result of interest (although its value may vary with, among other things, the range of stimulus intensities employed in any particular investigation). The same scale of loudness, adopted by the International Organization for Standardization, is based on this approach. Magnitude estimation has been profitably applied in the assessment of how one's degree of stage fright is related to the size of a prospective audience and of the analgesic value of hypnosis. It has also been used to evaluate the diminution of fear experienced by phobic persons over the course of therapy for their condition.

A procedure that should be equivalent to magnitude estimation is magnitude production, in which people are given a number and told to adjust a stimulus so that its attribute level is equal to that number. For example, a person might be told to adjust the intensity of a sound so that its loudness was 60. Results of this procedure also exhibit power function relations between numbers and stimulus intensities. However, the exponent relating numbers to stimulus intensities is rarely the same as that found with magnitude estimation and is usually larger. Where sufficient data are available, averaging the results of the two methods is often recommended for deriving a more accurate function relating stimulus intensities and sensation magnitudes.

The coherence of magnitude scaling approaches is seen in the procedure of cross-modality matching, in which, for example, people hear a sound and are asked to adjust a light's intensity so that its brightness is equivalent to the loudness of the sound. If one knows the two power functions relating loudness to sound intensity and brightness to light intensity, then one can predict the power function that will relate

sound intensity to light intensity at perceptual equivalence. Agreement of experimental data with these predictions is very good.

Although doubts remain as to the ultimate validity of these approaches, magnitude scaling methods have a central place in the scaling of perceptual magnitudes.

Category Scaling

More familiar are category scaling procedures, in which people are asked to rate things using a scale that ranges from 1 to 10 or from -5 to $+5$ or over some other range of numbers. Here again, the average ratings constitute the psychological scale values for the things being rated. Scale values for the same set of objects may vary in complicated ways depending not only on the range of stimuli employed but also on the particular rating scale used by the observers—both its range and its location. For example, when people were asked to rate how successful they had been in their lives, 34% of them chose a rating in the lower half of a -5 to $+5$ scale, but only 13% did so with a 1 to 11 scale, highlighting the fact that these scales do not provide information on the absolute levels of people's perceptions. In addition, people exhibit a tendency to try to use all available ratings roughly equally often on category scales; this can give distorted results if the things being rated have a skewed distribution on the attribute being rated.

Category-based scales and magnitude-based scales on the same stimuli are often found to be nonlinearly related, a problem that has not been satisfactorily resolved in general.

Difference-Based Scaling

Based on work of Roger Shepard and J. B. Kruskal, nonmetric scaling permits the construction of an interval scale using people's assessments of differences between objects. An exhaustive ranking of the sizes of the pairwise differences among objects can give rise to an interval scale for the values of the objects on the dimensions along which they vary (and to the number of dimensions, or attributes, being measured);

multidimensional differences can be handled with this approach. If there are n objects, then there are $n(n - 1)/2$ pairwise differences between them. Having 10 objects that vary in one dimension provides sufficient constraint to adduce interval-scale values for the 10 objects. This approach has been used successfully in, for example, the measurement of occupational prestige.

Item Response Theory

Georg Rasch, building on the work of Thurstone, developed item response theory (IRT), a scheme for measuring attributes such as knowledge of some domain in a way that can give rise to interval scales. IRT, unlike many other schemes, does not treat all items on a test as equivalent (and hence does not simply rely on a total score for its measure) but recognizes that the questions vary in difficulty so that there are some questions that only the most knowledgeable test taker will answer correctly and others that most people will answer correctly. Easy questions allow us to distinguish among the knowledge levels of people who know little because some such people will get more easy questions right than others will; easy questions cannot help us distinguish among the very knowledgeable, who will get all of them right. Similarly, hard questions will be gotten wrong by all those who know little but can help us distinguish among those who know a lot. The probability that someone will get a question right grows with that person's knowledge level. We can imagine, then, for any question, a graph relating the probability of a correct answer on the ordinate to the test taker's knowledge level on the abscissa. For easy questions, the probability of a correct answer rises to near 1.0 "early" in the graph, that is, at low knowledge levels. For a hard question, the probability will remain at zero until knowledge level is fairly high, at which point the probability will start to rise toward 1.0. Wherever along the abscissa the curve rises steeply defines a range of knowledge levels that the question helps us distinguish among; where the curve is flat or shallow, the question provides little discrimination among knowledge levels. IRT uses these ideas to produce simultaneous measurement of (a) question difficulty

and (b) knowledge levels, utilizing as input the question-by-question performance of a large number of testees. A person's pattern of correct and incorrect answers permits measurement of the person's knowledge level. In many cases, these are interval-scale measures.

—Scott Parker

See also Item Response Theory; Thurstone Scales

Further Reading

- Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.
- Dunn-Rankin, P., Knezek, G. A., Wallace, S., & Zhang, S. (2004). *Scaling methods* (2nd ed.). Mahwah, NJ: Erlbaum.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage.
- Marks, L. E., & Gescheider, G. A. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 91–138). New York: Wiley.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Mahwah, NJ: Erlbaum.
- Moskowitz, H. R. (1983). *Product testing and sensory evaluation of foods: Marketing and R&D approaches*. Westport, CT: Food & Nutrition Press.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- van der Ven, A. H. G. S. (1980). *Introduction to scaling*. Chichester, UK: Wiley.

SCAN STATISTIC

Scan statistics are used for detecting unusual clusters of points scattered randomly on a line or events occurring randomly over time. Traditionally, the scan statistic is defined as the maximum number of points contained in a window of fixed length w sliding along a continuous interval of real numbers, say, from 0 to 1. The points on the interval may represent events over

time. An exceptionally large value for the scan statistic will indicate the presence of a cluster.

In 1965, Joseph Naus showed that the scan statistic is the test statistic in a generalized likelihood ratio test for the null hypothesis of n points independently sampled from a uniform distribution on $[0,1]$ versus an alternative of the existence of a cluster. Naus proved that among a class of tests for the presence of clusters among uniformly distributed points, N_w is the most powerful test statistic. Since then, there have been numerous applications of scan statistics in mathematics, science, and industry, such as the generalized birthday problems, clustering of diseases in time, clustering of defective items in manufacturing processes, and many more.

How large a value must the scan statistic of a cluster have before it can be declared unlikely to occur by chance? To answer this question, one needs a mathematical formula for the probability distribution of the scan statistic. Finding the distribution function of the scan statistic has become an object of intense interest. In 1977, F. K. Hwang derived the general formula for the distribution of the scan statistic. However, the practical use of this elegant formula is severely restricted because of the formidable computations involved in the complicated sum of matrix determinants. Considerable efforts, therefore, have been channeled to finding accurate approximate distributions that are computationally more tractable. A number of these approximations are presented in the book *Scan Statistics and Applications* by J. Glaz and Balakrishnan, published in 1999.

In applying the scan statistics, choosing an appropriate value of the window size w is usually not a straightforward matter. In many cases, significant clusters may not be detected because the window size is either too small or too big. It is generally advisable to perform the analysis using a variety of window sizes in any application of scan statistics to detect clusters. It has been noticed that the repetitive applications of scan statistics on the computer can be carried out more conveniently if we use an equivalent form of scan statistics known as r -scan, put forth by Amir Dembo and Samuel Karlin at Stanford University in 1992, while the original form of scan

statistics is called w -scan. We present the definitions of both the w -scan and the r -scan below with an explanation of their duality relationship.

Let $X_1, X_2, \dots, X_n$ be n points independently sampled from the unit interval of real numbers between 0 and 1, and let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be their order statistics from the smallest to the largest. That is, $X_{(1)}$ is the smallest one among $X_1, X_2, \dots, X_n$, $X_{(2)}$ the second smallest, and so on. Let

$$S_i = X_{(i+1)} - X_{(i)}, i = 1, \dots, n-1$$

denote the spacing between adjacent points, and let

$$A_r(i) = S_i + \dots + S_{i+r-1}$$

be the sum of the r adjoining spacings starting at $X_{(i)}$. Also let $N_w(i)$ stand for the number of points contained in a window of length w beginning at $X_{(i)}$.

For a fixed window length $0 < w < 1$, the traditional scan statistic is the largest number of points contained in a window of size w scanning over the unit interval. Mathematically, it can be defined as $N_w = \max\{N_w(i): i = 1, \dots, n\}$. For a fixed positive integer r , the r -scan statistic is $A_r = \min\{A_r(i): i = 1, \dots, n-r\}$. In other words, the r -scan statistic A_r is the smallest of the aggregated r -spacings of the n points.

It is not hard to see that if there is a window of length w containing $r+1$ points or more, then there must be r adjoining spacings whose sum is less than w . The converse of this is also true. So N_w and A_r are related by a duality relation

$$\{N_w \geq r+1\} = \{A_r \leq w\}.$$

Consequently, if the distribution function of either one is known, so will the other. The traditional w -scan statistics and the new r -scan statistics can be used interchangeably.

An Application to Molecular Biology

In the early 1990s, advances in biotechnology generated an exponential growth of biomolecular sequence data. Scan statistics have played critical roles in

analyzing these data and extracting useful biological information. They have helped predict important functional sites in the genome and scan markers for disease susceptibility genes. We shall illustrate how the scan statistics were used in a specific example.

DNA is deoxyribonucleic acid, which consists of four different nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). A large number of these bases are strung together to form a giant sequence with a backbone composed of deoxyribose (a sugar) and phosphate groups. The nucleotide bases A and T form a complementary pair, as do C and G. Because of this pairing, the DNA molecule forms a double helix of two strands with complementary base sequences.

Viral genomes are small, each of them comprising a DNA molecule ranging from a few thousand to several hundred thousand bases. Early studies in molecular virology have reported that the nucleotide sequences around replication origins of certain herpesviruses have a high concentration of palindromes. A DNA palindrome can be defined as a word pattern of the form $b_1 \dots b_L b'_L \dots b'_1$, where b' denotes the complement of base b and L is the half-length of the palindrome. For example, the sequence of bases GCAATATTGC is a palindrome of length 10.

As the central step in the reproduction of herpesviruses, viral DNA replication has been the target for a number of antiherpesvirus drugs (e.g., acyclovir). Understanding the molecular mechanisms involved in DNA replication is of great importance in further developing strategies to control the growth and spread of viruses. Since replication origins are regarded as major sites for regulating genome replication, labor-intensive laboratory procedures have been used to search for replication origins. With the increasing availability of genomic DNA sequence data, one way that may save time and resources would be to scan the viral genome sequence for the expected sequence features by a computer program before an experimental search for replication origins is launched.

The human cytomegalovirus (HCMV) is a member of the herpesvirus family, which includes several well-known viruses, such as herpes simplex and

chicken pox. The HCMV genome is a DNA molecule of 229,354 base pairs. The computerized search for replication origins began by scanning through the genome sequence for palindromes of length 10 or more and located 296 such palindromes. Idealizing the occurrences of these palindromes on the genome sequence as 296 points randomly scattered on the interval of numbers between 0 and 1, scientists have made calculations based on the scan statistics that indicate the presence of a statistically significant (p value < 0.01) cluster of palindromes in the segment from 92,001 to 93,000 bases. Highly sensitive experimental assays performed around this part of the genome confirmed that the segment between 92,210 and 93,175 bases is the OriLyt of HCMV.

In 1994, Leung and coworkers first attempted to formalize the mathematical ideas behind the successful prediction of the OriLyt in HCMV using scan statistics to identify unusual clusters of palindromes and confirmed the feasibility of using palindrome clusters for predicting replication origins in general for the herpesvirus family. In 2005, based on the scan statistics with compound Poisson approximation, Leung and coworkers reported the details of the mathematical theory, suggesting possible ways to improve the reliability of their prediction method of replication origins.

The Assumptions Underlying the Scan Statistics

In the definition of scan statistics, one has to assume that a certain number n of points are sampled independently from the interval of real numbers between 0 and 1 according to a uniform distribution. That is, each point is selected independently and equally likely from anywhere in the interval. It is generally advisable to make a probability plot of the data points against the uniform quantiles to verify that this assumption is satisfied overall to a reasonable degree.

The Research Hypothesis

The scan statistics are used when the researcher suspects that among the observed data points, there is an

unusual cluster that is unlikely to occur by chance. In the test of clustering by the scan statistics, the null hypothesis assumes that all the points are independent and identically distributed according to a uniform distribution. Assuming that the null hypothesis is true, the probability (p value) of observing a value of the scan statistics as high as the value in the data is assessed. If the p value is too small, say, less than 0.05, the null hypothesis will be rejected, and we shall then say that a nonrandom cluster exists among the points.

Computations Involved

Computations involved in the scan statistics are quite complicated. Many different programs based on different approximations of the probability distribution of the scan statistics have been developed. (See, for example, the ones described in *Scan Statistics and Applications*, by Glaz and Balakrishnan.) To fit the different needs of specific applications, it is quite common for researchers to develop their own computer codes. For example, Josephine Hoh and Jurg Ott developed their Scanstat program for gene mapping, available online. Other application codes not online can usually be obtained from the authors directly.

—Ming-Ying Leung

Further Reading

- Dembo, A., & Karlin, S. (1992). Poisson approximations for r -scan processes. *Annals of Applied Probability*, 2(2), 329–357.
- Glaz, J., & Balakrishnan, N. (1999). *Scan statistics and applications*. Boston: Birkhauser.
- Hoh, J., & Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences USA*, 97(17), 9615–9617.
- Hwang, F. K. (1977). A generalization of the Karlin-McGregor theorem on coincidence probabilities and an application to clustering. *Annals of Probability*, 5(5), 814–817.
- Leung, M. Y., Choi, K. P., Xia, A., & Chen, L. H. Y. (2005). Nonrandom clusters of palindromes in herpesvirus genomes. *Journal of Computational Biology*, 12(3), 331–354.
- Leung, M.-Y., Schachtel, G. A., & Yu, H.-S. (1994). Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. *Nonlinear World*, 1, 445–471.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60, 532–538.

Scanstat program instructions, from J. Hoh & J. Ott: <http://linkage.rockefeller.edu/ott/Scanstat.html>

SCATTERGRAM

A scattergram is a visual summary demonstrating the relation between two variables; it illustrates the scores on one variable plotted against scores on another variable. The scattergram is usually drawn before working out a linear correlation coefficient or fitting a regression line. It is used as an aid to interpret the data. For instance, the scattergram can illustrate whether there are any outliers or gaps in the data or whether the data follow a straight or a curved line.

The pattern obtained from the scattergram illustrates the type and strength of the relation between two variables. If the data points make an upward trend going from bottom left to top right (positive slope), then the association is positive (the relation is direct; as one variable increases, so does the other). If the straight line goes down from top left to bottom right (negative slope), then the association is negative (the relation is indirect; as one variable increases, the other decreases). The closer the points follow a straight line, the higher the correlation. No visible trend from the scattergram implies no association; the points are scattered randomly. If the data points form a curved line, then a linear correlation cannot be used to describe the data.

In Figure 1, the scattergram illustrates a negative relation between age and traffic tickets. The scattergram was created using Excel.

When creating a scattergram, observe the following conventions:

1. There are no gridlines for the X or Y axis.
2. The axes are clearly labeled.
3. The points are not connected by a line.

—Adelheid A. M. Nicol

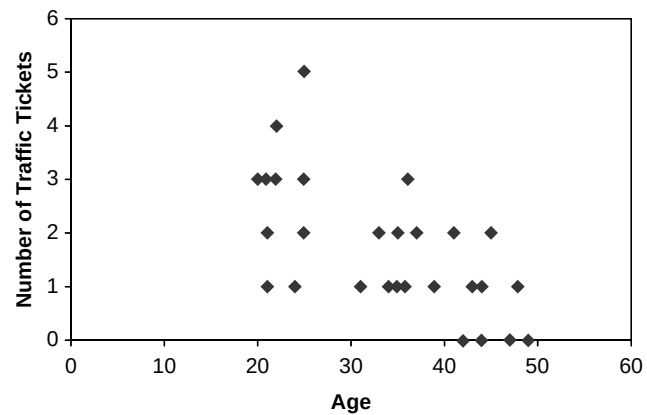


Figure 1 Total Number of Traffic Tickets Issued to Individuals Age 20 to 50 in 2005

See also Correlation Coefficient; Histogram; Line Chart; Regression Analysis

Further Reading

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41, 103–130.

Wilkinson, L. (1994). Less is more: Two- and three-dimensional graphics for data display. *Behavior Research Methods, Instruments, & Computers*, 26, 172–176.

SCREE PLOT

In an explorative data analysis, principal component analysis and factor analysis are often applied for the purposes of data reduction or structure detection or simplification. In these analyses, it is important to decide the number of components or factors to be retained. While there is no set rule, there are a few common conventions. One common practice is to consider only those factors having variance of factors (or eigenvalues) greater than 1. This method is referred to as the *eigenvalue criterion method*. The other common practice is to use a scree plot. In this case, the plot and the eigenvalue criterion method are often used in combination.

A scree plot is a two-dimensional plot with the number of factors on the horizontal axis and eigenvalues on the vertical axis. Its purpose is to provide a useful visual aid for determining an appropriate number of components to retain. The plot shows the fraction of total variance in the data as explained, or represented, by each factor or component. The plot is called a scree plot because it looks like a sloping mass of loose rocks (scree) at the base of a cliff. The scree plot allows one to pick the number of components or factors on the basis of the point at which “elbow,” or separation, is observed or where a plateau begins. The plot in Figure 1 shows eigenvalues versus number of factors with annotation of the first five values of the eigenvalues and their corresponding cumulative percent of variance. The eigenvalues are based on the results from Proc Factor in SAS, and the plot is created by Minitab software.

In this plot, an elbow occurs at the third point and is followed by a plateau. That is, the eigenvalues after the third value are all relatively small. Thus, we see that three principal components should be retained for further use. There is a clear separation between the first two components and the remaining components.

—Kyoungah See

See also Eigenvalues; Factor Analysis; Factor Scores

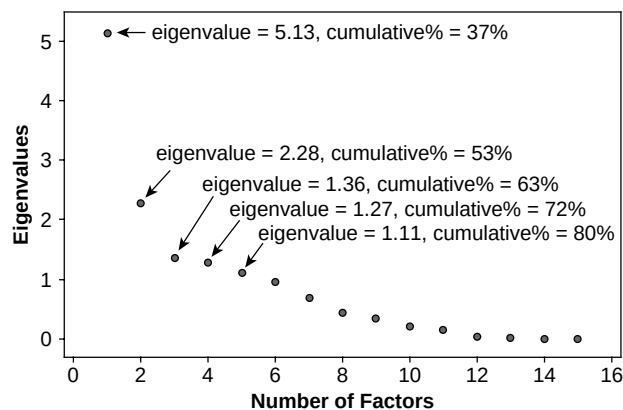


Figure 1 Scree for Nursing Home Quality Data

Further Reading

Johnson, R. A., & Wichern, D. (2002). *Applied multivariate statistical analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

Exploratory factor analysis primer: http://www.upa.pdx.edu/IOA/newsom/semclass/ho_efa.doc

Scree plots: <http://www.rsd-associates.com/mtxscree.htm>

SECONDARY DATA ANALYSIS

If locating a sample on which to conduct research poses problems, it might be worthwhile to consider conducting a secondary data analysis. During the past 25 years, an increasing number of researchers in the United States have turned to secondary data rather than collecting original data for their research. More recently, some researchers have begun to explore uses of secondary data for qualitative research.

At its most basic level, secondary data analysis involves using data collected in previous research to address a different research question. Special considerations in any secondary analysis include differences in operational definitions, unavailability of raw data, and various ethical issues. This entry explores these issues as well as the uses, advantages, and disadvantages of secondary data analysis. To demonstrate how secondary data analysis has been used to address various topics, summaries of some secondary analysis studies have been included at the end of this entry.

Similarities and Distinctions Between Secondary Data Analysis and Meta-Analysis

Graves considered meta-analysis to be a special case of secondary analysis. However, because of differences in the two methods, some researchers do not agree with Graves. A comparison of these two methods reveals both similarities and distinctions. Regarding similarities, both methodologies use data collected by others. Both also offer the temptation to

rely on published findings, which means that each method can be subject to publication bias.

However, at least two distinctions exist between secondary analysis and meta-analysis. First, secondary data analysis uses previously collected data to address problems other than those for which the data were collected. In contrast, meta-analysis combines results from multiple studies with similar hypotheses to clarify findings on the same problem. Second, whereas meta-analysis entails using data from multiple studies, secondary data analysis might use data from a single (usually large) study. Because of these differences between the two methods, some researchers have viewed secondary analysis as distinct from all other types of research.

Practical Basis for Secondary Data Analysis

Governments, businesses, schools, and other organizations have gathered much quantitative information. Therefore, when available information gathered for primary research can serve other research purposes, it has made sense to use that information, rather than repeatedly attempting massive data collection efforts. Researchers have highlighted at least four practical reasons for using secondary data.

Efficiency

Secondary data has been less costly and frequently more easily available than primary data. This has been confirmed by a variety of researchers. Other secondary data analysis scholars concurred that, with research funding shortages, secondary data banks have provided a cost-efficient source of data.

Strengthened Confidence

Both Brannigan and Cowton noted that similar trends found by different sources strengthen confidence in findings. This phenomenon, known as triangulation, often can be achieved efficiently through multiple secondary sources. For triangulation purposes, K. J. Kiecolt and L. E. Nathan also

recommended combining original research with analysis of secondary data.

Special Need for Some Studies

The textual data found in newspaper accounts, letters, and other qualitative sources has been essential for historical research. Katz (in “How satisfied are the self-employed: A secondary analysis approach”; see the last research example in this entry) made the point that secondary data might be less biased than some types of primary data, such as self-report interviews. The reason proposed for reduced bias was that self-report is more subjective than third-party reporting.

Another example of using qualitative data to meet special research needs is found in Lindenmann’s award-winning public health study. Since public health information often is presented to general audiences in qualitative form, Lindenmann chose newspaper articles as the venue for his study. To learn how newspapers reported on matters affecting public health, he studied two national and five regional newspapers’ articles that related to public health issues.

Protection of Research Pool

Finally, secondary data analysis might have particular application for graduate students. Fitzpatrick claimed that by using the analysis from secondary information while learning to do research, potential research participants are protected.

Research Uses of Secondary Data Analysis

Quantitative researchers have found broad applications for secondary analysis because of the availability of summative data in national data banks. These data can help in addressing a host of educational, organizational, medical, environmental, and social problems. Also, as noted elsewhere in this article, graphic and textual secondary data have become widely available to address qualitative research questions. Qualitative data banks include those found in periodicals, newspapers, and other print media.

Quantitative Uses

Trochim and Van Dierendonck, Schaufeli, and Buunk noted that multiple secondary sources were used for some research questions. For example, Trochim noted that a researcher could study crime patterns in different parts of the United States using both census and crime data; likewise, a study reported by Van Dierendonck and colleagues noted use of five sources of secondary data for their examination of causal relationships among the three dimensions of burnout. Similarly, census data could be combined with discipline-specific databases to examine relevant issues in other fields. Some recent examples of uses of quantitative secondary analysis are studies of job satisfaction, work environment, gender, and worker decision making.

Qualitative Uses

Historical studies, obviously, have depended on secondary sources. Common sources for secondary data have included ethnographies, legislation, treaties, annual reports of various agencies, and letters and diaries. These sources have been used to study such topics as trends in social behavior; effects of substance abuse; health care; business; and work, school, and family issues. Also, mixed-methods studies have used both qualitative and quantitative as well as primary and secondary data.

Secondary Data Analysis Process

The process for secondary data analysis, as noted in Figure 1, is similar to that for other types of research except for the methods of gathering and working with the data. For original research, researchers (a) formulate research questions, (b) conduct literature reviews, (c) develop research designs, (d) collect data, (e) analyze data, and (f) report findings. For secondary analyses, they (a) formulate research questions, (b) conduct literature reviews, (c) establish criteria for inclusion, (d) locate summarized data collected by others *or* use their own primary data for a different research purpose, (e) analyze data, and (f) report findings.

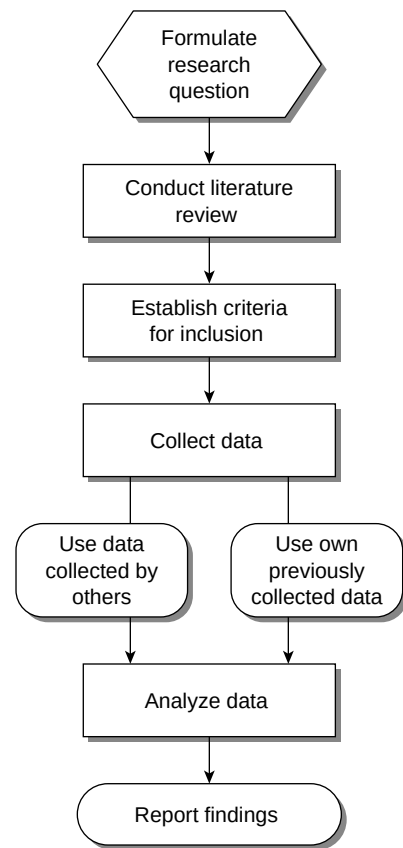


Figure 1 Steps in Quantitative Secondary Data Analysis

In the data gathering and recording stages of secondary data analysis, researchers establish size of study, research design, quality controls, and so forth, for the data to be included. Next, they search data archives for data meeting their criteria. These data, frequently subsets of larger data sets, are transported into SPSS or another statistical program. Variables are established, often by combining other variables or by constructing new variables. Once the secondary data have been analyzed, secondary researchers compare their findings with findings of primary researchers and report their findings.

Variations of the Basic Secondary Data Analysis Process

There have been some variations, or modifications, of the basic secondary data analysis process. These process changes have arisen from the unique needs of

research projects. A few of these modifications are discussed in this section, including (a) use of secondary and primary data in the same study, (b) secondary data analysis using one's own data, and (c) studies that include both qualitative and quantitative secondary data.

Using secondary data analysis to complement primary data studies. Heaton ("Children and Families Team") used a combination of primary and secondary data. This approach can strengthen confidence in the findings of secondary data analysis while adding depth to answers to research questions.

Secondary data analysis of researcher's own data. Researchers can also reuse their own data for a new research purpose. Using this variation, researchers could avoid some of the challenges ordinarily involved in secondary data analysis because they would understand the primary research fully and would have access to the complete data set.

Studies using both qualitative and quantitative secondary data. Finally, Janet Heaton, in "Secondary analysis of qualitative data," discussed the option of including both qualitative and quantitative secondary data in studies. For example, a researcher might use secondary data analysis to identify aspects of job satisfaction in middle-management positions in the accounting industry. The researcher might then conduct interviews with several middle managers in selected accounting firms to add another dimension to the study. This variation can reinforce and clarify the findings of primary studies.

Major National Archives

Several large databases have been developed and made available to researchers. These may be classified according to research disciplines or by agency; some examples are included in this discussion. However, Mertens advised that some national databases have limitations relating to ways they can be manipulated. For example, a researcher might find it impossible to disaggregate data according to demographic variables such as gender or race. Also, some demographics that

interested a researcher—such as type of disability, educational level, and so forth—might not even be included in the study.

Nongovernment Databases

ISI Web of Knowledge

Developed by ISI (<http://www.isinet.com/>), this site was introduced in 2001. ISI, which is owned by the Thomson Corporation, has provided resources to researchers since the 1950s. It has become a worldwide organization with offices in the United States, Ireland, Japan, and China. The site contains qualitative and quantitative data banks on a variety of scientific, business, and social topics, and information is available in several languages.

Shuji Kaneko, an employee of Thomson, believes that a major advantage of data on the Web is its direct connection to related information through the link function. ISI, he said, can create links from secondary databases to primary sources. Kaneko also believes that ISI used adequate critical selection criteria to provide reliable and useful information. The criteria he listed were authority, accuracy, currency, navigation and design, applicability and content, scope, audience level, and quality of writing.

General Social Survey

The General Social Survey was one of three databases for social science research recommended by researchers. The General Social Survey is updated every year or so through major national surveys conducted through the University of Chicago. The purpose of the surveys has been to make fresh data available to researchers. Data has included international information as well as that collected in the United States.

Poll Service Databases

The Roper Center for Public Opinion Research, with facilities located at the University of Connecticut, has conducted public opinion polls on a variety of socially relevant issues, including information from the United States as well as 70 other nations. Other well-known opinion poll services include the Gallup

Poll and national news service polls. While sharing agreements for these data vary, interested researchers could find helpful information through these and other poll services.

Commercial Databases

In “Secondary analysis in entrepreneurship: An introduction to databases and data management,” Katz recommended marketing databases for research in appropriate areas. Some of those included Dun and Bradstreet, TriNet, and BRS. Katz also recommended that researchers consider the Interuniversity Consortium for Political and Social Research (<http://www.icpsr.umich.edu/>) Guide to Resources and Services for data-sharing opportunities.

Hard Copy Data Sources

According to Lawrence Neuman, excellent hard copy sources for social and political research include the *Almanac of American Politics*, *America Votes: A Handbook of Contemporary American Election Statistics*, and *Vital Statistics on American Politics*. These sources are available to researchers through their national congressional representatives. The *Almanac of American Politics* publishes congressional voting records, *America Votes* has county voting information for most statewide and national offices, and *Vital Statistics* provides information on campaign spending and a variety of related issues. Information gathered from these databases could be used to amend current practices or to develop policy for new initiatives.

Government Databases

Unlike some of the databases mentioned earlier, government databases can be accessed without charge. Another advantage of government data has been that it may be less biased than other data sources. A description of some government databases follows.

The Data Archive

This is a United Kingdom site. It is a major storage for quantitative computer files and their documentation

in the United Kingdom, according to the University of Southampton. The Data Archive also contains qualitative material—primarily research reports in the broad areas of humanities and social sciences (i.e., education, labor, history, politics, household finance, child development, crime, and public health).

FedWorld

This site (<http://www.fedworld.gov/>) promises to make it easy to locate government information. It has links to countless resources for secondary data analysis. These resources include the census and education databases, discussed next.

Census Databases

The U.S. Bureau of the Census database could be useful to researchers in a variety of disciplines, including business, social science, politics, health, and education. The national census conducted at the beginning of each decade gathers information on population numbers, ages, educational and income levels, housing, employment, and other areas. Inferential statistics can be derived from the vast store of current and past information from the censuses made available through the U.S. Departments of Commerce, Education, and Labor as well as directly through the U.S. Bureau of the Census.

Survey of Income and Program Participation

The Survey of Income and Program Participation database is available through a collaborative effort between the U.S. Census Bureau and the U.S. Social Services Administration. This database contains frequently updated information on income, employment, and participation in various government financial assistance programs.

Education Databases

The National Center for Education Statistics has developed a special database that has been of great assistance to persons preparing grant proposals for educational programs. This database can be found through the U.S. Department of Education Web site

(at <http://www.ed.gov/index.jhtml>). It provides information on demographics according to state, age, level of education, and so forth.

Another valuable resource for education research is the database for the Higher Education Research Institute (<http://www.gseis.ucla.edu/heri/heri.html>). It has information available at a minimal charge.

Advantages of Secondary Data Analysis

Various authors, such as Katz (in “Secondary analysis in entrepreneurship: An introduction to databases and data management”) and Trochim, have discussed advantages of secondary data analysis. This section contains a discussion of some of the most notable reasons for those advantages. Those reasons are efficiency, conservation of participant pools, and the advantage of using unobtrusive measures to obtain the data.

Efficiency

Some authors, including Jean Barclay, A. Dale and colleagues, E. J. Graves, R. Katz (in “Secondary analysis in entrepreneurship: An introduction to databases and data management”), and Kiecolt and Nathan, have stated that conservation of time, effort, and money is the strongest reason for secondary analysis. They pointed out that, especially for studies involving very large populations, individual researchers did not have the necessary money or access to obtain data firsthand. For example, Katz recommended secondary analysis for business studies involving entrepreneurial issues.

Conservation of Participant Pools

J. Fitzpatrick noted that large numbers of graduate students could “wear down” significant pools of potential research participants. Wimmer and Dominick noted that although novice researchers, such as students, can gain benefits from developing and conducting research, this particular type of analysis doesn’t usually produce results that are externally valid.

Unobtrusive Measures

Several researchers noted that, since secondary data analysis involves data collection in an indirect manner, they could avoid some contamination of the data. This observation might apply especially to quantitative data. However, it also could invite some of the challenges of secondary analysis because the burden of selecting high-quality studies would become increasingly important.

Challenges of Secondary Data Analysis

In addition to being a very useful tool, secondary analysis presents challenges to the researcher.

Getting a good fit. Accessing data collected for one research purpose and using it for a different purpose require carefully thought-out criteria for selection to assure the validity of the secondary analysis study. Also, there are limitations on the kinds of research questions and hypotheses that can be used for secondary data analysis purposes. A researcher must adapt the study to questions that can be answered using available data, and the primary study or studies must meet criteria for responsible research.

Loss of control of data collection. Obviously, when a researcher has not been involved in developing the research design or in the collection of data, a loss of control over the process of collecting the data necessarily occurs. In addition, the secondary analyst usually has access only to summarized data.

Ethical issues. Heaton (in “Secondary analysis of qualitative data”) noted that ethical issues could become important in secondary analysis. For example, should participants be contacted for permission to use their information for a different research purpose? Would the new study involve any compromise of either the identity of participants or the integrity of the data?

Establishing new variables. To meet the objectives of a new research purpose, a researcher often must

manipulate the secondary data. This could involve taking subsets of data from databases. It also could require a researcher to collapse information into new variables.

Research Applications of Secondary Data Analysis

Internationally, researchers have used secondary data to study education, business, and social issues. This section contains a review of human resource studies, including decision making at the middle management level, job satisfaction among volunteers, and job satisfaction among entrepreneurs.

Managerial Decision Making

An applied ethics study by Harris examined relationships between courage and middle management decision making. Harris used secondary qualitative data from 610 newspaper items he found in four leading international newspapers. Specifically, the data were examples of courageous behavior and situations in which the virtue of courage was needed for making management decisions. He stated he used reporter accounts because he believed that the data sources were less biased than self-reporting data would have been.

Harris conducted a content analysis of the gathered data. The process involved him in (a) selecting newspaper reports that provided the necessary data, (b) using special software to code the data, and (c) analyzing and evaluating the data. Harris also conducted a check for face validity using the phrases he had coded, noting that the percentage, reliability coefficient, and agreement coefficient were calculated. Reliability coefficients ranged from 77% to 97%; agreement coefficients ranged from 70% to 90%. According to Harris, his validity and reliability results were consistent with or superior to those of other researchers on the variable of courage.

Findings from this study indicated that accounts of courage could be recognized in the descriptions of courage and courageous activity in major newspapers throughout the world. However, Harris found

insufficient data to provide useful information relative to differences in professions or particular ethical values or choices.

Job Satisfaction of Volunteers

Silverberg and his colleagues studied job satisfaction among volunteers in the public parks and recreation industry. The purpose of their study was to check the reliability and the validity of the questionnaire used in a primary study of almost 6,000 volunteers who responded to a “modified employee job satisfaction scale.” Silverberg and colleagues noted that 50% of American adults are volunteers in non-profit organizations and that volunteer effectiveness depended on competence of volunteers and low turnover.

They determined that the instrument was a reliable, valid instrument for making inferences about volunteer job satisfaction in the parks and recreation industry. They also found that volunteer satisfaction was related to job setting and to the psychological needs met through the volunteer job function. Specifically, volunteers in this study wanted to increase their understanding of governmental and organizational operations. Working to benefit others was not seen as a prime motivator.

Recommendations from the study by Silverberg and colleagues included (a) instituting regular meetings between managers and volunteers to gain the feedback necessary to ensure high job satisfaction among volunteers and (b) trying to ensure a good match between volunteer assignments and volunteer motives.

Job Satisfaction Among Entrepreneurs

In “How satisfied are the self-employed: A secondary analysis approach,” Katz used secondary analysis of three previous studies involving more than 1,200 participants to study job satisfaction among the self-employed. He concluded that self-employed workers had higher job satisfaction rates than did salaried workers. Katz also noted that having access to secondary data was insufficient for secondary analysis. Also necessary was detailed information on

(a) how the primary researcher handled missing data, (b) the procedures followed in sample and subsample selection, and (c) coding variations that would affect analysis. Katz recommended that journal editors allow more space so that researchers could report the details needed for secondary analysis. He also suggested that there be changes in research practices in the area of entrepreneurship to enable other researchers to use their data to best effect.

—Ernest W. Brewer

Further Reading

- Cowton, C. J. (1998). The use of secondary data in business ethics research. *Journal of Business Ethics*, 17, 423–434.
- Harris, H. (2001). Content analysis of secondary data: A study of courage in managerial decision-making. *Journal of Business Ethics*, 34(3/4), 191–208.
- Katz, J. A. (1992). Secondary analysis in entrepreneurship: An introduction to databases and data management. *Journal of Small Business Management*, 30(2), 74–75.
- Silverberg, K. E., Marshall, E. K., & Ellis, G. D. (2001). Measuring job satisfaction of volunteers in public parks and recreation. *Journal of Park and Recreation Administration*, 19(1), 79–92.
- Wimmer, R. D., & Dominick, J. R. (2000). *Mass media research: An introduction*. Boston: Wadsworth.

SECTION 504 OF THE REHABILITATION ACT OF 1973

Section 504 of the Rehabilitation Act of 1973, which traces its origins to the wake of World War I, at which time the U.S. government sought to provide vocational rehabilitation to injured soldiers, is the oldest federal law addressing the needs of the disabled. According to Section 504, in language that is similar to that in Title VI of the Civil Rights Act of 1964 and Title IX of the Educational Amendments of 1972: “[n]o otherwise qualified individual with a disability in the United States . . . shall, solely by reason of her or his disability, be excluded from the participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving [f]ederal financial assistance . . .” (29 U.S.C.A. § 794(a)).

Section 504 differs significantly from the Individuals with Disabilities Education Act (IDEA). For example, while Section 504 applies to school systems that receive federal financial assistance in the form of money, books, or free lunches, school systems do not receive additional funds under its provisions, as do boards that serve children who qualify for IDEA services. Further, Section 504 protects individuals under the broader notion of impairment rather than the IDEA’s reliance on the statutorily concept of disability, and Section 504 has no age limitation. Moreover, Section 504 covers students, employees, and others, including parents, while the IDEA focuses on the rights of children.

Section 504 defines an individual with a disability as one “who (i) has a physical or mental impairment which substantially limits one or more of such person’s major life activities, (ii) has a record of such an impairment, or (iii) is regarded as having such an impairment” (29 U.S.C.A. § 706(7)(B)). In order to have “a record of impairment,” one must have a history of, or have been identified as having, a mental or physical impairment that substantially limits one or more major life activities, including learning and working (45 C.F.R. § 84.3(j)(2)(i)). Once an individual is identified as having an impairment, educators must consider whether the child is “otherwise qualified.” Unlike the IDEA, Section 504 neither requires that individuals submit to psychological or other examinations nor affords extensive due process protections.

In order to be “otherwise qualified,” persons must be “(i) of an age during which nonhandicapped persons are provided such services, (ii) of any age during which it is mandatory under state law to provide such services to handicapped persons, or (iii) [one] to whom a state is required to provide a free appropriate public education [under the IDEA]” (45 C.F.R. § 84.3(k)(2)). Individuals who are “otherwise qualified” (meaning that, impairment aside, they are eligible to participate in programs or activities) must be permitted to partake as long as they can be provided with “reasonable accommodations.”

Even if one appears to be “otherwise qualified,” educators can rely on three defenses to avoid being charged

with noncompliance with Section 504 (another major difference between Section 504 and the IDEA, which does not recognize defenses). First, educators can be excused from making accommodations that would result in “a fundamental alteration in the nature of [a] program” (*Southeastern Community College v. Davis*). The second defense allows educators to avoid making a modification if it imposes an “undue financial burden” (*Southeastern Community College*). The third defense is that otherwise qualified individuals can be excluded from programs if their presence creates a substantial risk of injury to themselves or others (*School Bd. of Nassau County v. Arline*).

—Charles J. Russo

Further Reading

Osborne, A. G., & Russo, C. J. (2006). *Special education and the law: A guide for practitioners* (2nd ed.). Thousand Oaks, CA: Corwin.

Public welfare. *Code of Federal Regulations*, Title 45.

School Bd. of Nassau County v. Arline, 480 U.S. 273, 287–88 (1987).

Section 504 of the Rehabilitation Act of 1973, 29 U.S.C.A. § 794(a).

Southeastern Community College v. Davis, 442 U.S. 397 (1979).

Title VI of the Civil Rights Act of 1964, 42 U.S.C.A. §§ 2000 *et seq.*

Title IX of the Education Amendments of 1972, 20 U.S.C.A. § 1681.

Council for Exceptional Children: <http://www.cec.sped.org>
 Disabilities Education Improvement Act (IDEA) of 2004
 resources: <http://www.ed.gov/offices/OSERS/IDEA/>
 Office of Special Education Programs: <http://www.ed.gov/about/offices/list/osers/osep/index.html?src=mr>

SELF-REPORT

Self-report is one of the most commonly utilized methods of data collection. Surveys, questionnaires, and interviews are all forms of self-report that rely on individuals' subjective evaluations and reports of their thoughts, feelings, behaviors, or experiences. Most often, self-report is used to gather personal

information that cannot be obtained objectively. An individual's self-report may also be of interest in circumstances in which some degree of objective evaluation is possible (e.g., a patient's subjective reports of symptoms). In many settings, such as medicine, policy making, and opinion polls, important decisions are made on the basis of self-report data.

There are two broad categories of self-report data: (a) unstructured, open-ended responses, and (b) structured, fixed-response questions. Open-ended reports have the advantage of soliciting detailed information that may not be captured using closed-response formats. In many cases, this approach is optimal because it provides a wealth of qualitative data. However, such responses do not easily lend themselves to statistical analysis without the use of coding procedures, which can be labor intensive and difficult to develop.

Closed-response formats generate quantitative data much more easily. However, decisions regarding the use of checklists, number scales, or categorical endorsements can have a dramatic impact on response distributions and the subsequent presence or absence of significant statistical effects. Research has demonstrated that by simply altering the response format of a questionnaire, one can generate different responses to the same questions. For instance, decisions regarding whether to anchor a response scale with a midpoint, whether to use a bipolar (–5 to +5) or unipolar (0 to 10) number scale, and whether to label responses with verbal quantifiers like “frequently” or “very often” can influence one's perception of the questions and subsequently the responses one endorses.

A number of cognitive constraints can also interfere with one's ability to generate accurate self-reports. Both the time scale and the regularity of a behavior can alter the memory retrieval strategy used to recall events, while the accessibility of recent experiences may influence more general estimates. Similarly, measures designed to evaluate emotional traits or memories are often biased by the influence of current mood states. Research on such contextual effects has also demonstrated that responses to self-report items may be directly influenced by the content and placement of previous self-report items (i.e., later

answers are influenced by earlier answers). Furthermore, respondents' conscious beliefs regarding the confidentiality of their reports can affect their disclosure of personally sensitive information (especially regarding threatening or stigmatized topics).

Critically evaluating questions to ensure that they are presented clearly, framed in the proper context, and accompanied by appropriate response formats can help prevent self-report data from being compromised by measurement constraints or response biases. Clearly informing respondents as to the intended use, privacy, and protection of self-report information can also reduce self-presentation concerns and facilitate more veridical reporting. A number of innovative self-report methodologies, such as daily diaries and ecological momentary assessment, have addressed some of these concerns by considerably limiting the recall periods (i.e., to a day or even a few minutes) and providing an ecologically valid alternative to lengthy retrospective reporting. By carefully considering these issues, researchers can effectively use self-report as a fast, cheap, and practical method for collecting personal information across a variety of research and applied settings.

—Joshua Smyth and Christopher P. Terry

See also Measurement; Personality Tests

Further Reading

- McLaughlin, M. E. (1999). Controlling method effects in self-report instruments. *Research Methods Forum*, 4. Retrieved May 17, 2006, from http://www.aom.pace.edu/rmd/1999_RMD_Forum_Method_Effects_in_Self-Reports.htm
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.

SEMANTIC DIFFERENTIAL

Semantic differential is a term used to refer to a method of measuring different dimensions of meaning underlying responses toward an object. This procedure involves presenting individuals with opposing adjective pairs (e.g., good-bad, strong-weak, fast-slow) and asking them to identify where on the scale they feel the object fits in relation to the two adjectives.

Semantic differential scales can be used to capture three broad dimensions of meaning underlying reactions to an object: evaluation, potency, and activity. However, in practice, most researchers use the procedure to assess global attitudes toward an object. Thus, in most situations, researchers focus exclusively on the evaluation dimension.

Development of the Semantic Differential Method

The origins of the semantic differential method can be traced to the pioneering work of Charles Osgood and his colleagues. Osgood was interested in understanding the meaning that people attached to words. He noted that words have both a denotative meaning and a connotative meaning. Denotative meaning refers to the literal dictionary definition of a word. Connotative meaning, on the other hand, refers to the associations of meaning that are attached to a word that are not, strictly speaking, part of its formal definition.

Osgood was interested in identifying the basic dimensions of connotative meaning underlying words. To explore this issue, he constructed pairs of opposing adjectives and then had respondents rate objects on these adjective pairs. Using factor analyses, he found that adjectives seemed to reflect one or more of the three basic underlying dimensions of connotative meaning mentioned above: evaluation, potency, and activity. Evaluation refers to the good-bad continuum of meaning underlying words. Potency reflects the strong-weak continuum. Activity represents the active-passive dimension. These dimensions of meaning have consistently been replicated, and analyses by Osgood and subsequent researchers have identified certain adjective pairs as highly representative of the evaluation (e.g., good-bad, valuable-worthless), potency (e.g., large-small, strong-weak), and activity (e.g., fast-slow, hot-cold) dimensions.

Constructing Semantic Differential Scales

Drawing on this research on connotative meaning, Osgood and others have proposed that semantic

Item selection also requires the researcher to keep in mind that adjective pairs must be easily related to the object being rated. Some adjective pairs are quite generic and are likely applicable to nearly any object. For example, good-bad could probably be sensibly applied to almost any physical object, person, social group, or concept. Other adjective pairs may reflect somewhat more specific meaning and thus be less easily applied to some objects than others. For instance, beautiful-ugly is highly evaluative in content and could quite sensibly be applied to a person (e.g., one's spouse) and many physical objects (e.g., a car, a work of art). This pair might be less meaningfully applied to other objects or concepts (e.g., magnetic resonance imaging, social security).

When presenting semantic differential items to assess multiple objects, several strategies can be used. First, an object may be presented and then followed with the items seeking to measure that object. For example, DOG may be presented at the top of a page, and underneath it are the descriptive items such as

good-bad, like-dislike, and others. Alternatively, the object and one item that the object is to be rated on can be presented. The object then changes but the item remains the same. For example, DOG is presented with a good-bad scale. This item is followed by CAT, which is also presented with a good-bad scale, and so on. In yet another method of presentation, objects being rated on

[illegible]

Figure 1 Sample of Semantic Differential Scale Measuring Attitudes Toward Vegetarianism

Check the position between each adjective pair that best describes the meaning of *vegetarianism* to you.

Bad	:	_____												:	Good
Bad		1	–	2	–	3	–	4	–	5	–	6	–	7	Good
Bad		–3		–2		–1		0		+1		+2		+3	Good

Figure 2 Sample of Semantic Differential Response Formats

different adjective pairs may be presented randomly. Research suggests that results do not vary considerably between these methods, although the first method of presentation allows for easier organization and coding of data.

Once responses to items have been collected, an overall score for ratings of the object on the dimension of interest is computed. For instance, when scoring a semantic differential scale assessing evaluation, researchers usually code responses from 1 to 7 or from –3 to +3, with higher numbers reflecting more positive evaluations. Responses to individual items are then summed or averaged to provide an overall attitude score. Scoring is similar for scales assessing potency and activity, with higher numbers typically reflecting greater potency or activity.

Strengths and Weaknesses of Semantic Differential

The greatest strengths of the semantic differential method are its ease and versatility. Because previous analyses have identified adjective pairs that are highly representative of the dimensions of evaluation, potency, and activity, it is reasonably easy to select items to construct a semantic differential scale. Moreover, many adjective pairs are quite general and thus can be broadly applied to a wide range of objects (e.g., people, groups, concepts, and physical objects). These features make constructing a semantic differential scale a relatively easy process. They also permit reasonably direct comparisons of attitudes toward different objects because the same set of items can be used to assess attitudes toward the objects being compared. Yet another strength of the semantic differential scales is that they have been found to perform well with as few as four items. This efficiency permits quick measurement of attitudes and makes it feasible to

assess attitudes toward numerous topics in a short period of time.

In contrast, other well-known attitude scaling techniques are more cumbersome and less flexible. For example, some of these methods (e.g., Thurstone scales and Guttman scales) require significant item pretesting and evaluation.

These alternative methods also typically produce items that are attitude object specific (i.e., items are applicable to only a single target of judgment, such as a specific social issue, a particular person, or single product). Hence, these alternative methods often require more time and effort for developing a scale, and the scale is usually useful only for a single attitude object or a narrow class of attitude objects. Some of these alternative methods also typically require more items for assessing attitudes. For instance, Thurstone and Likert scales usually involve no less than 10 items and generally as many as 20 or 30 items. Given the practical advantages of the semantic differential method, it is not surprising that semantic differential scales are perhaps the most popular method of attitude measurement.

Formal evaluations of the semantic differential method have suggested that it is a reasonably effective method of attitude measurement. Semantic differential scales are routinely highly correlated with attitude measures constructed using other scaling techniques. Moreover, because of the popularity of the method, semantic differential scales have been used extensively in basic and applied attitude research. In these contexts, they have been found to be sufficiently sensitive for detecting experimental manipulations designed to alter attitudes and for predicting outcomes postulated to be influenced by attitudes.

Although semantic differential scales have proven useful, methodologists also point out potential limitations. For example, respondents who read each item in a strictly literal sense may miss out on the implied meaning within the adjective pairing. For example, if the object were SELF and the item included dirty-clean, the respondent might be confused and not know how to answer. To help remedy this potential problem, participants are often told to go with their first instinct. In addition, careful selection of adjective

pairs can help ensure that very general items are used or that the specific items used are clearly applicable to the object being judged.

A second potential problem is that semantic differential scales are fairly obvious in their objectives. Thus, respondents may figure out what is being measured and may report their attitudes in a self-enhancing light. Of course, such a limitation is true of nearly all self-report measures and can sometimes be reduced with assurances of anonymity.

Another proposed limitation of the method is that semantic differentials allow only for overall analysis of the evaluative dimension of meaning. However, some researchers have explored the possibility of identifying word adjective pairs that are highly affective or cognitive in nature in addition to being highly evaluative. This research suggests that constructing more-specific semantic differential measures of affect and cognition (and perhaps other specific dimensions) is possible.

Finally, it is important to note that the bipolar structure of semantic differentials does present some ambiguity in interpreting scores. Specifically, mid-point responses can reflect one of several reactions to the object. It could imply that a respondent is ambivalent toward the object (i.e., the respondent has strong conflicting positive and negative evaluations), that the respondent truly is neutral (i.e., the respondent's reactions are mostly nonevaluative), or that the respondent has no opinion.

—Leandre R. Fabrigar and Meghan E. Norris

See also Attitude Tests; Guttman Scaling; Likert Scaling; Thurstone Scales

Further Reading

- Crites, S. L., Jr., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, 20, 619–634.
- Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In C. T. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 17–40). Thousand Oaks, CA: Sage.
- Heise, D. R. (1970). The semantic differential and attitude research. In G. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Fort Worth, TX: Harcourt Brace Jovanovich.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teacher's College Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Connotative meaning information: <http://www.writing.ws/reference/history.htm>

SEMANTIC DIFFERENTIAL SCALE

A semantic differential scale is an efficient method of measuring various dimensions of meaning underlying responses toward an object. This procedure involves presenting individuals with opposing adjective pairs and asking them to locate the object on a rating scale anchored by the opposing adjectives. Semantic differential scales can be used to capture three broad dimensions of meaning underlying reactions to an object: evaluation, potency, and activity. Evaluation represents the good-bad continuum with respect to an object, potency reflects the strong-weak continuum, and activity represents the active-passive continuum. Although the method can be used to assess all three of these dimensions, in practice many researchers choose to measure only the evaluation dimension. When the scale consists only of evaluation items, it can be conceptualized as a measure of attitudes. Indeed, semantic differentials are among the most widely used methods of attitude measurement in the social sciences.

A semantic differential scale usually consists of 4 to 10 items (i.e., adjective pairs). Each descriptive item contains two adjectives, opposite in meaning, on either end of a continuous scale that typically has seven levels of response. Participants are asked to identify where on the scale they feel the object fits in relation to the two adjectives. For example, the object may be DOG, and items might include good-bad and valuable-worthless. Participants then respond to each item by indicating where on the continuum they would place the object in terms of direction and intensity, with a neutral response being in the middle of the scale.

Table 1 Sample of Semantic Differential Evaluation Items

<i>Vegetarianism</i>														
Dislike	1	–	2	–	3	–	4	–	5	–	6	–	7	Like
Good	1	–	2	–	3	–	4	–	5	–	6	–	7	Bad
Wise	1	–	2	–	3	–	4	–	5	–	6	–	7	Foolish
Pleasant	1	–	2	–	3	–	4	–	5	–	6	–	7	Unpleasant

When scoring a semantic differential scale assessing evaluation, responses are coded from 1 to 7 or from –3 to +3, with higher numbers reflecting more positive evaluations. Responses to individual items are then summed or averaged to provide an overall score. Scoring is similar for scales assessing potency and activity.

When creating a semantic differential scale, researchers should ensure that appropriate adjective pairs have been selected to assess the dimensions of interest. Published factor analyses of word meanings can help determine which dimensions adjective pairs best reflect. In addition, researchers need to consider the appropriateness of word pairs for the specific object of interest. Some adjective pairs are quite general and are probably appropriate for virtually any object. For example, good-bad could probably be meaningfully applied to virtually any physical object, person, social group, or abstract idea. Other adjective pairs may be somewhat more specific and thus perhaps less readily applied to some objects than others. For instance, beautiful-ugly could quite sensibly be applied to a person but less meaningfully applied to an object such as magnetic resonance imaging.

—*Leandre R. Fabrigar*
and *Meghan E. Norris*

See also Semantic Differential

Further Reading

Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In C. T. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 17–40). Thousand Oaks, CA: Sage.

Heise, D. R. (1970). The semantic differential and attitude research. In G. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Fort Worth, TX: Harcourt Brace Jovanovich.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teacher's College Press.

SEMI-INTERQUARTILE RANGE

Briefly, the semi-interquartile range is a measure of the dispersion or spread of a variable; it is the distance between the 1st quartile and the 3rd quartile, halved.

It is common to describe a variable using a measure of central tendency, or average, most commonly the mean or median. However, in order to make sense of a measure of average, we need to have a measure of dispersion. When the mean is used as a measure of average, the standard deviation is usually used as the measure of dispersion. When the median is used, it is more appropriate to use the semi-interquartile range.

The semi-interquartile range is preferred over the range because it is not affected by extreme scores. The range is calculated using only two data points: the highest and the lowest. If one of these values were to change, the range would change dramatically.

**Example of the Calculation
of the Semi-Interquartile Range**

Table 1 shows the data contained in one variable, sorted in order of magnitude and numbered, to make life easier.

There are 21 values, and the middle point, which is the median, is the middle one. This is the 11th point and is equal to 9. To find the exact point, add 1 to the sample size and multiply by 0.5. Thus, the median

Table 1 Finding the Quartiles of the Distribution

1.	0	
2.	1	
3.	2	
4.	3	
5.	4	
1st quartile, (25th centile) = 4.5		
6.	5	
7.	5	
8.	8	
9.	9	
10.	9	
	9	Median (2nd quartile, 50th centile) = 9
11.	12	
12.	12	
13.	12	
14.	13	
15.	13	
3rd quartile (75th centile) = 13.5		
16.	14	
17.	14	
18.	15	
19.	18	
20.	22	

point is given by $(21 + 1) \times 0.5 = 11$; it is the 11th point. There are 10 people who score higher than this and 10 who score lower.

For the quartiles, we want the points that are one quarter and three quarters of the way through the distribution. We use a procedure similar to the one we used for the mean, except multiplying by .25 and .75. The points are therefore obtained as follows:

$$\begin{aligned}\text{lower quartile} &= (21 + 1) \times .25 = 5.5; \\ \text{upper quartile} &= (21 + 1) \times .75 = 16.5.\end{aligned}$$

The obvious problem we encounter is that these values do not exist. To solve this problem, as we would with the median, we take the mean of the nearest two values, weighted by their distance from the value that we want. For the 1st quartile, we want the value in the 5.5th position. We therefore find the mean of the fifth and sixth values. The mean of 4 and 5 is 4.5, so the 1st quartile is 4.5. Similarly, for the 3rd quartile, we want the value that is in 16.5th place, and

so we use the mean of the 16th and 17th values. The mean of 13 and 14 is 13.5.

Finally, the semi-interquartile range is the difference between these two values, halved:

$$(13.5 - 4.5) / 2 = 4.5.$$

Additional Notes

The calculation of the quartiles is complicated when the values for the quartiles do not land exactly in between two potential values. For example, if we have a sample of size 32, the 1st quartile will be the 8.25th position. If the 8th value is 4 and the 9th value is 5, the interquartile range should be closer to the 8th value than the 9th and hence will be 4.25, not 4.5.

In a symmetrical distribution, the range given by the median plus or minus the semi-interquartile range will include approximately half of the values. In our case, this would be 11 ± 4.5 , giving a range of 6.5 to 15.5. In our data set, this range includes 13, slightly more than half. (The semi-interquartile range is used instead of the interquartile range because of this property.)

When the data are normally distributed, the sampling distribution of the semi-interquartile range is wider than the sampling distribution of the standard deviation, and hence it is not used when the standard deviation is appropriate.

—Jeremy Miles

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

Semi-interquartile range page: <http://davidmlane.com/hyperstat/A48607.html>

SHAPIRO-WILK TEST FOR NORMALITY

The most popular nongraphical procedure for testing for fit to the normal distribution is the Shapiro-Wilk test. The test can be obtained easily from leading statistical packages such as R, SAS, and SPSS. This is

fortunate because the Shapiro-Wilk test statistic W is laborious to calculate by hand. The Shapiro-Wilk test statistic is obtained by dividing the square of an appropriate linear combination of the sample order statistics by the sum of squares error. The formula is

$$W = \frac{\left\{ \sum_{i=1}^n a_i (x_{(n-i+1):n} - x_{i:n}) \right\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where a_i represents special coefficients obtained from a table in Shapiro and Wilk's 1965 report or a source such as Conover. A computer algorithm given by Royston will also approximate these coefficients. The formulas are also found online in the Wikipedia. After W is calculated, the hypothesis of normality is rejected if W is less than a quantile from a value in another special table.

One major flaw in the original form of the Shapiro-Wilk test was that the table with the necessary coefficients and percentage points was available only for samples of size 50 or less. To deal with this flaw, Shapiro and Francia adapted the W statistic so that the necessary coefficients for the linear combinations depended only on the expected values of the normal order statistics, which are more readily available. Another table of empirically obtained percentage points was given. Royston extended the Shapiro-Wilk test for sample sizes up to 2,000. Royston's method involves an approximate normalizing transformation on the W statistic, which is not asymptotically normal. The calculations are tedious by hand but easily programmed into a computer. In 1989, Royston provided a correction for cases in which ties are present in the data. Royston's version of the Shapiro-Wilk test is available on SAS.

In the Shapiro-Wilk test, the null hypothesis is defined to be that the data are normally distributed (with some unspecified mean and standard deviation), and the alternative hypothesis is that the data are not normal. Therefore, a rejection of the null hypothesis, which will occur when the value of the W statistic is small and the p value is less than the specified level of significance (usually $\alpha = .05$), indicates a significant deviation from normality. A value of W close to its

maximum of 1 indicates a close fit to normality, which is indicated by a p value that is above $\alpha = .05$, and a failure to reject the null.

Unfortunately, the Shapiro-Wilk test shares many of the standard flaws of most goodness-of-fit tests. When the sample is small, the test lacks power, and often the test will fail to reject a sample that arises from a non-normal population. This is troublesome because it is these small sample situations in which most analysts are the most concerned with testing for normality. The usual reason to perform a goodness-of-fit test for normality is to determine whether the assumption of normality is met well enough for the analyst to perform standard parametric tests (e.g., t test, ANOVA). If the assumption is not met, then the researcher may need to make a transformation or pursue alternatives such as nonparametrics, bootstrapping, or Bayesian methods. Also, the Shapiro-Wilk test will almost always reject its null hypothesis when the sample is very large. With a data set of several thousands, even minor deviations present in data generated from a known normal distribution will lead to rejection of normality.

Example

To avoid extreme tedium, we will rely on a computer package to perform the calculations. The examples given here utilize the `shapiro.test` function in the free statistical package R. The Shapiro-Wilk test is also available in virtually all standard statistical software packages, such as SAS and SPSS. SAS users can obtain the Shapiro-Wilk test by adding the normal option to PROC UNIVARIATE.

The first example uses a data set of size 100 randomly generated from a normal distribution of size 100 with mean 100 and standard deviation 15. In this case, $W = 0.9889$ and the p value = 0.5775. We do not have enough evidence to reject normality. Figure 1 is a density plot (or essentially a smoothed histogram) of these data.

The second example (Figure 2) uses a data set of size 100 randomly generated from an exponential distribution of size 100 with mean 1. In this case, $W = 0.8646$, and p value = 4.254×10^{-8} . Notice that the smaller value of the W statistic leads to a p value so small that we will reject the null hypothesis for any

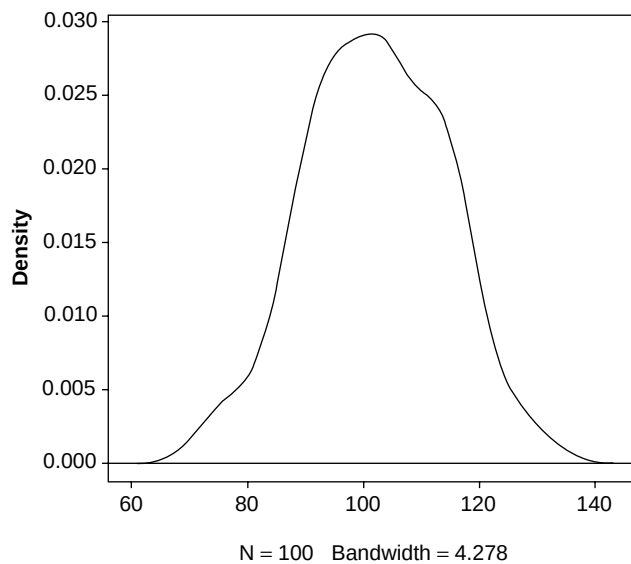


Figure 1 Density (x = Normal)

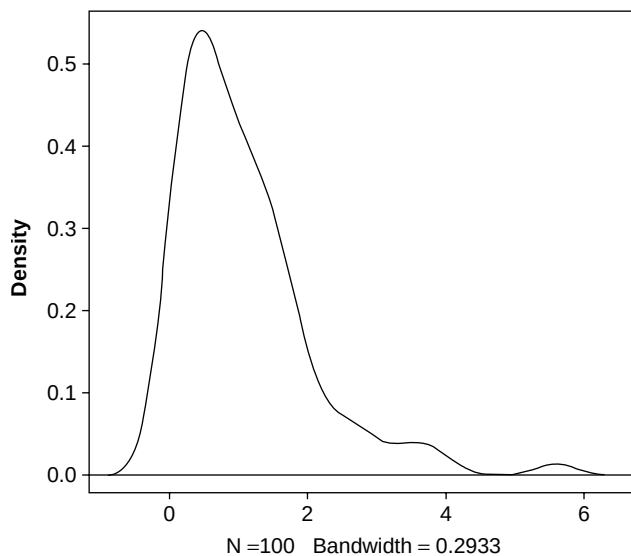


Figure 2 Density (x = Exponential)

reasonable alpha level and therefore conclude that these data are not normally distributed.

Power

The power of the Shapiro-Wilk test has been compared with several competitors, such as the chi-square, Cramer-von Mises, and Kolmogorov-Smirnov tests and tests based on the standardized third moment (skewness) and fourth moment (kurtosis). It was found to be generally superior to the other tests across a

wide range of possible alternatives. The empirical distribution function tests such as Kolmogorov-Smirnov and Cramer-von Mises, which are based on distance, had lower power. A combination of the skewness and kurtosis tests had more power but was still usually outperformed by the Shapiro-Wilk W statistic.

Recommendations for Testing Univariate Normality

D'Agostino, Belanger, and D'Agostino recommended that the skewness and kurtosis tests should always be calculated as descriptive statistics and that the tests based on $\sqrt{b_1}$ and b_2 should be used, along with Q-Q or normal probability plots. Pearson, D'Agostino, and Bowman compared the $\sqrt{b_1}$ test, the b_2 test, the Shapiro-Wilk W test, the Shapiro-Francia W' test, D'Agostino's Y test, and the D'Agostino-Pearson K^2 test against dozens of alternative distributions. While no test is uniformly the most powerful against all possible distributions, the W test of Shapiro and Wilk was generally the best performer. D'Agostino noted that while the Shapiro-Wilk test is a very sensitive omnibus test against skewed alternatives, one may wish to use a directional test, such as the $\sqrt{b_1}$ test, if the direction of deviation from normality in terms of skewness is known a priori. The b_2 test is more powerful than Shapiro-Wilk's W or D'Agostino's Y if the direction of deviation from normality in terms of kurtosis is known. D'Agostino recommended against the use of chi-square or Kolmogorov-Smirnov tests because of their low power.

—Christopher J. Mecklin

See also Lilliefors Test for Normality

Further Reading

- Conover, W. G. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.
- D'Agostino, R. B. (1986). Tests for the normal distribution. In R.B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 367–419). New York: Marcel Dekker.
- D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44, 316–321.
- Pearson, E. S., D'Agostino, R. B., & Bowman, K. O. (1977). Tests for departures from normality: Comparisons of powers. *Biometrika*, 64, 231–246.

- Royston, J. P. (1982). Expected normal order statistics (exact and approximate). *Applied Statistics*, 31, 161–165.
- Royston, J. P. (1982). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics*, 31, 115–124.
- Royston, J. P. (1982). The *W* test of normality. *Applied Statistics*, 31, 176–180.
- Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67, 215–216.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, 52, 591–611.
- Shapiro, S. S., Wilk, M. B., & Chen, H. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63, 1343–1372.

Shapiro-Wilk test: http://en.wikipedia.org/wiki/Shapiro-Wilk_test

SIGNAL DETECTION THEORY

Signal detection theory (SDT) is used to analyze data coming from experiments in which the task is to categorize ambiguous stimuli that can either be generated by a known process (in which case the stimuli are called the *signal*) or be obtained by chance (in which case they are called the *noise* in the SDT framework). For example, radar operators must decide whether what they see on a radar screen indicates the presence of a plane (the signal) or the presence of parasites (the noise). This type of application was the original framework of SDT (see the founding work by Green & Swets). But the notion of signal and noise can be somewhat metaphorical in some experimental contexts. For example, in a memory recognition experiment, participants have to decide whether the stimulus they currently see was presented before. Here the signal corresponds to a familiarity feeling generated by a memorized stimulus whereas the noise corresponds to a familiarity feeling generated by a new stimulus.

The goal of detection theory is to estimate two main parameters from the experimental data. The first parameter, called d' , indicates the strength of the signal (relative to the noise). The second parameter, called C (a variant of which is called β), reflects the strategy of response of the participant (e.g., saying easily Yes

rather than No). SDT is used in disparate domains, from psychology (psychophysics, perception, memory) to medical diagnostics (do the symptoms match a known diagnostic or can they be dismissed as irrelevant?) to statistical decision (do the data indicate that the experiment has an effect or not?).

The Model

It is easier to introduce the model with an example, so suppose we have designed a face memory experiment. In the first part of the experiment, a participant is asked to memorize a list of faces. At test, the participant is presented with a set of faces one at a time. Some faces in the test were seen before (these are *old* faces) and some were not seen before (these are *new* faces). The task is to decide for each face whether this face was seen (response Yes) or not (response No) in the first part of the experiment.

What are the different types of responses? A Yes response given to an old stimulus is a correct response, and it is called a *Hit*, but a Yes response to a new stimulus is a mistake, and it is called a *False Alarm* (abbreviated as FA). A No response given to a new stimulus is a correct response called a *Correct Rejection*, but a No response to an old stimulus is a mistake, called a *Miss*. These four types of response (and their frequency) can be organized as shown in Table 1.

The relative frequency of these four types of response is not all independent. For example, when the signal is present (first row of Table 1), the proportion of Hits and the proportion of Misses add up to 1 (because when the signal is present, the participant can say either Yes or No). Likewise, when the signal is absent, the proportion of FAs and the proportion of Correct Rejections add up to 1. Therefore, all the information in a table such as Table 1 is given by the proportion of Hits and FAs.

Table 1 The Four Possible Types of Response in Signal Detection Theory

Reality	Decision (Participant's Response)	
	Yes	No
Signal present	Hit	Miss
Signal absent	False Alarm	Correct Rejection

Even though the proportions of Hits and FAs provide all the information in the data, these values are hard to interpret because they crucially depend on two parameters. The first parameter is the difficulty of the task: The easier the task, the larger the proportion of Hits and the smaller the proportion of FAs. When the task is easy, we say that the signal and the noise are well separated or that there is a large distance between the signal and the noise. (Conversely, for a hard task, the signal and the noise are close, and the distance between them is small.) The second parameter is the strategy of the participant: A participant who always says No will never commit an FA (and will never get any Hit either); on the other hand, a participant who always says Yes is guaranteed all Hits (but will also have all FAs). A participant who tends to give the response Yes is called liberal, and a participant who tends to give the response No is called conservative.

The SDT Model

So the proportions of Hits and FAs reflect the effect of two underlying parameters: the first one reflects the separation between the signal and the noise, and the second one the strategy of the participant. The goal of SDT is to estimate the value of these two parameters from the experimental data. In order to do so, SDT creates a model of the participant's response. The SDT model assumes that the participant's response depends on the intensity of a hidden variable (e.g., familiarity of a face) and that the participant responds Yes when the value of this variable for the stimulus is larger than a predefined threshold.

SDT also assumes that the stimuli generated by the noise condition vary naturally for that hidden variable. As is often the case elsewhere, SDT, in addition, assumes that the hidden variable values for the noise follow a normal distribution. Recall at this point that when a variable x follows a Gaussian (also called normal) distribution, this distribution depends on two parameters: the mean (denoted μ) and the variance (denoted σ^2). It is defined as follows:

$$\mathcal{G}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}. \quad (1)$$

In general, within the SDT framework, the values of μ and σ are arbitrary, and therefore we choose the simpler values of $\mu = 0$ and $\sigma = 1$ (other values will give the same results but with more cumbersome procedures). In this case, Equation 1 reduces to

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}x^2 \right\}. \quad (2)$$

Finally, SDT assumes that the signal is *added* to the noise. In other words, the distribution of the values generated by the signal condition has the same shape (and therefore the same variance) as the noise distribution.

Figure 1 illustrates the SDT model. The x -axis shows the intensity of the underlying hidden variable (e.g., familiarity for the face example). As indicated above, the distribution of the noise is centered at zero (i.e., mean of the noise is equal to zero, with a standard deviation of 1). So the standard deviation of the noise is equivalent to the unit of measurement of x . The distribution of the signal is identical to the noise distribution, but it is moved to the right of the noise distribution. The distance between the signal and the noise distributions corresponds to the effect of the signal (this is the quantity that is added to the noise distribution in order to get the signal distribution): This distance is called d' . Because the mean of the noise distribution is 0, d' is equal to the mean of the signal distribution.

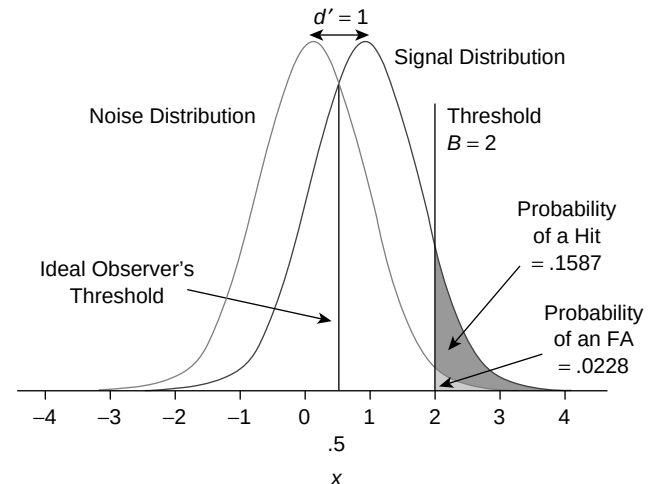


Figure 1 The Model of SDT

The strategy of the participant is expressed via the choice of the threshold. There are several ways of expressing the position of this threshold; among the possible candidates, we will mention four of them, denoted B , D , C , and β . The first quantity, B (sometimes called θ), gives the position of the threshold on the x -axis. In the example illustrated in Figure 1, this value is equal to 2, and so the participant corresponding to this figure has decided that any stimulus with a value of x larger than 2 comes from the signal distribution and is given the response Yes. The position of the threshold can also be given relative to the signal distribution (because the noise has zero mean, B is the distance of the threshold relative to the noise distribution); as the mean of the signal is equal to d' , we can compute D as $D = d' - B$ (a value equal to 1 in our example).

The most popular way of expressing the location of the threshold, however, is from neither the distribution of the noise nor the distribution of the signal but relative to what is called the ideal observer. The ideal observer minimizes conjointly the probability of a Miss and of an FA. When each type of error has the same cost, the criterion of the ideal observer is positioned on the average of the means of the signal and the noise distribution. In our example, the threshold of the ideal observer would be equal to $\frac{1}{2}d' = \frac{1}{2} = .5$. The value of C is the distance from the actual threshold to the ideal observer; it can be computed as $C = B - \frac{d'}{2} = 2 - .5 = 1.5$. The sign of C reveals the participant's strategy: when $C = 0$, we have the ideal observer; when C is negative, the participant is liberal (i.e., responds Yes more often than the ideal observer); when C is positive, the participant is conservative (i.e., responds No more often than the ideal observer).

An alternative way of expressing the position of the participant's criterion is given by the quantity called β . It corresponds to the ratio of the height of the signal distribution to the noise distribution for the value of the threshold. Because the distributions of the noise and the signal are normal with variance equal to one, we can compute β from Equation 2 as follows:

$$\beta = \frac{\mathcal{N}(D)}{\mathcal{N}(B)} = \frac{\mathcal{N}(1)}{\mathcal{N}(2)} = \frac{.2420}{.0540} \approx 4.4817. \quad (3)$$

Equation 3 can be rewritten as follows:

$$\beta = \exp\{d' \times C\}. \quad (4)$$

The quantity β has the advantage of being a *likelihood ratio* and can be used to interpret SDT within a statistical framework. For practical reasons, it is often easier to compute the logarithm of β ; for example, from Equation 4, we get

$$\ln \beta = d' \times C = 1 \times 1.5 = 1.5. \quad (5)$$

The model illustrated by Figure 1 generates a specific pattern of response probabilities which can be computed from integrating the normal distribution. So, for example, the probability of an FA is obtained as the probability (i.e., area under the normal distribution) of finding a value larger than 2 with a normal distribution of mean 0 and variance 1 (this can be computed with most statistical packages or from tables such as the ones given in Abdi). This quantity is also called the probability associated with the value 2; in our example, it is equal to .0228. Along the same lines, the probability of a Hit is obtained as the probability (i.e., area under the normal distribution) of finding a value larger than 2 with a normal distribution of mean 1 (i.e., the mean of the signal) and variance 1; this is equivalent to finding the probability (i.e., area under the normal distribution) of finding a value larger than $2 - 1 = 1$ with a normal distribution of mean

Table 2 The Probability of the Four Possible Types of Response According to Figure 1

Reality	Decision (Participant's Response)		Total
	Yes	No	
Signal present	Hit	Miss	1
	$Pr \{HIT\} = .1587$	$Pr \{Miss\} = .8413$	
Signal absent	False Alarm (FA)	Correct Rejection	1
	$Pr \{FA\} = .0228$	$Pr \{Correct Rejection\} = .9772$	

$1 - 1 = 0$ and variance 1. This value is equal to .1587.

SDT in Practice

The previous example described the performance of a participant who behaved according to the SDT model. However, in practice we do not know the values of the parameters of SDT, but we want to estimate them from the performance of the participants. In an experimental paradigm, the only observable quantities are the participant's responses, from which we can derive the number of hits and FAs.

To illustrate this problem, suppose we want to evaluate the performance of a wine taster whose task is to detect whether a wine labeled as made from Pinot Noir has been tempered by the addition of some Gamay (generally considered an inferior grape). Here the signal corresponds to the presence of Gamay. Our wine taster tasted (blindfolded) 20 glasses of Pinot, half of them tempered with some Gamay and half without. The results are reported in Table 3 and show that the proportions of Hits and FAs are .9 and .2, respectively. In order to find the values of d' and the criterion, we need to inverse the formulas given above (i.e., Equations 3–5). We also need one new notation: For a normal distribution with zero mean, we denote by Z_p the value of the normal distribution whose associated probability is equal to P (e.g., $Z_{.025} = 1.96$). We denote Z_H and Z_{FA} the values corresponding to the proportions of Hits and FAs. With these new notations and after some (minor) algebraic manipulations, we find the following set of formulas. The estimation of d' is obtained as

$$d' = Z_H - Z_{FA} = Z_{.9} - Z_{.2} = 1.28 - (-.84) = 2.12. \quad (6)$$

The estimation of C is obtained as

$$\begin{aligned} C &= -\frac{1}{2}[Z_H + Z_{FA}] = -[Z_{.9} + Z_{.2}] \\ &= -\frac{1}{2}[1.28 - .84] = -.22, \end{aligned} \quad (7)$$

Table 3 The Performance of a Wine Taster Trying to Identify Gamay in a Pinot Noir Wine

Reality	Decision: (Taster's Response)		Σ
	Yes (Gamay)	No (pure Pinot)	
Signal present (Gamay)	Hit # {Hit} = 9 Pr {Hit} = .9	Miss # {Miss} = 1 # {Miss} = .1	10 1
Signal absent (Pure Pinot)	False Alarm (FA) # {FA} = 2 Pr {FA} = .2	Correct Rejection # {Correct Rejection} = 8 Pr {Correct Rejection} = .8	10 1

and $1n \beta$ is obtained as

$$1n \beta = d' \times C = 2.12 \times -.22 = -.47 \quad (8)$$

(β is obtained as $\exp\{1n \beta\} = .63$).

How to interpret these results? The taster is clearly (but not perfectly) discriminating between Pinots and tempered Pinots (as indicated by a d' of 2.12). This taster is also liberal (in case of doubt, the taster would rather say that the wine has been tempered than that it has not).

—Hervé Abdi

See also Discriminant Analysis; Discriminant Correspondence Analysis; Distance; z Scores

Further Reading

- Abdi, H. (1987). *Introduction au traitement des données expérimentales*. Grenoble, France: Presses Universitaires de Grenoble.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
- McNicol, D. (1972). *A primer of signal detection theory*. London: Allen & Unwin.
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford, UK: Oxford University Press.

SIGNIFICANCE LEVEL

An integral part of every quantitative research study is the need to determine an appropriate statistical

significance level, α , also referred to as p_{critical} . Before data are collected, this level should be selected (a priori) because this level is logical, and the information gleaned from rejecting the null hypothesis is meaningful. In medical studies, it is common to set the significance level to $p < .01$, whereas $p < .05$ is common in the social sciences. For comparison, a $p_{\text{calculated}}$ or *Test Statistic*_{calculated} is used. The decision is the same regardless of whether a $p_{\text{calculated}}$ or *Test Statistic*_{calculated} is used. The $p_{\text{calculated}}$ is now easily obtained from myriad software packages. These probabilities $p_{\text{calculated}}$ and α are interesting in and of themselves. First, what do these probabilities mean? What is the interpretation of $p_{\text{calculated}}$ or *Test Statistic*_{calculated}?

The Probabilities: Alpha and $p_{\text{calculated}}$ or *Test Statistic*_{calculated}

Researchers may correctly use statistical significance testing in two general cases: (a) they have a random sample from a population from which an inference is to be drawn, or (b) they believe their sample approximates a random sample. Once this decision is made, the next is to set an alpha level. The setting of the alpha level, a probability ranging between 0 and 1, can be interpreted as the percentage chance of making a sampling error. For instance, an alpha level set at .05 indicates that there is a 5% chance of making an incorrect inference because sampling error creeps into all data without exception, barring collecting data from the entire population. When picking an alpha level, we set a boundary on the probability of making this incorrect inference, called a Type I error. Therefore, alpha is typically set small so that the probability of this error will be low. Thus, this alpha level, also termed p_{critical} , is selected on the basis of judgment regarding Type I error consequences in any given research situation, guided by personal values regarding these consequences.

After the data are collected and analyzed, part of the output is the second probability, often termed $p_{\text{calculated}}$. From this point forward, I treat $p_{\text{calculated}}$ and *Test Statistic*_{calculated} as being the same because the probabilistic interpretations are similar, and the interpretations are exactly the same. This probability, like

the alpha level probability, ranges between 0 and 1 and is calculated based on study parameters. There are two absolutely essential aspects impacting the calculation of the probability $p_{\text{calculated}}$. First is the assumption that the true population parameters are correctly described in the null hypothesis. This assumption is necessary because the actual population parameters are not known. Therefore, we assume that the null hypothesis is true for all calculations. Second, as the sample size approaches the population size, sampling error decreases, and the statistics calculated from them become more representative. It is intuitive that larger samples are more representative. When we think about this in the opposite direction, small samples are more likely to be comprised of less representative data points, and the statistics calculated for them potentially will be less representative of the population. To illustrate this point, if one were to catch two fish from a pond and both were catfish, one might draw the conclusion that the pond contains only catfish. However, with a little more effort and a larger sample, say 100 fish caught, one might learn that the pond ecosystem also consists of bass, perch, gar, and pickerel. So sample size is accounted for in the $p_{\text{calculated}}$ computations. For example, Experiment 1 consists of two groups with 10 members each. The mean for Group 1 is 55 and the mean for Group 2 is 56. The $p_{\text{calculated}}$ will be large and probably exceed the a priori alpha level of $p < .05$. However, Experiment 2 also consists of two groups (Groups 3 and 4), but this time each group contains 110 members. The mean for Group 3 is exactly the same as Group 1, 55, and the mean for Group 4 is exactly the same as for Group 2, 56 (see Table 1). This time $p_{\text{calculated}}$ is probably small and likely to be less than the a priori alpha level $p < .05$. Assuming the sample data were randomly selected from the population where the null hypothesis is true, what is the probability of obtaining the sample statistics from the given sample size(s)? The question in both experiments is, "Are the two means statistically significantly different with $\alpha = .05$?" In Experiment 1, the $p_{\text{calculated}} = .571$, so the means are not statistically different, but in Experiment 2, the $p_{\text{calculated}} = .043$, so the means are statistically significantly different even though the difference in the

Table 1 Descriptive Statistics

	Group	Mean	N	SD
Experiment 1	1	55.0	10	5.3
	2	56.0	10	6.7
Experiment 2	3	55.0	110	5.3
	4	56.0	110	6.7

Note: N = group size; SD = standard deviation.

means in both experiments is only 1 point (see Table 2). However, the Cohen's d effect size does not change from Experiment 1 to Experiment 2. When $p_{\text{calculated}}$ is less than p_{critical} or alpha, we use a decision rule that says we will reject the null hypothesis. The decision to reject the null hypothesis is called a *statistically significant result*. All the decision means is that we believe our sample results are relatively unlikely, given our assumptions, including our assumption that the null hypothesis is exactly true.

Implications

Quantitative results are often misinterpreted because few applied researchers understand the purpose of statistical significance testing. The partial understandings of aspects of statistical significance testing can result in systemic issues that can negatively impact the quality of their research reports. For example, researchers may understand the issue of sample size but worry only about samples that are too small while not considering the implications of extremely large samples or the implications for interpretations based on test statistics of these large samples. It is essential to augment null hypothesis testing with measures of treatment effects.

Statistical Level, Statistical Significance Tests, and Meta-Analytic Thinking

Statistical significance reporting is still an essential component of the newest edition of the *Publication Manual of the American Psychological Association* and a legislative reality and will continue to be reported. However, it is important to understand that given a large enough sample, one would always achieve statistical significance. More than a decade ago, E. N. Pedhazur and L. Schmelkin stated that few methodological issues have generated as much controversy as statistical testing. The $p_{\text{calculated}}$ values in a given study result from specific study characteristics such as sample size and effect size. The result is different studies, composed of varying sample sizes and effect sizes, all possibly having the same $p_{\text{calculated}}$ value. Second, $p_{\text{calculated}}$ does not infer from the sample to the population but from the population to the sample. One possible explanation for the prevalence of these misconceptions is that textbooks and graduate courses have been less than ideal in their treatment of statistical significance testing and its alternatives.

—Robert M. Capraro

Further Reading

- Capraro, R. M., & Capraro, M. M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, 62, 771–782.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.

Table 2 Independent Samples t Test

	Groups	N	Mean Diff.	SD	T	df	p (two-tailed)	d
Experiment 1	1, 2	20	−1.0	5.4	−.588	9	.571	.166
Experiment 2	3, 4	220	−1.0	5.1	−2.048	109	.043	.166

Notes: N = group size; Mean Diff. = mean difference; SD = standard deviation; T = t test statistic; df = degree of freedom; p = associated probability; d = Cohen's d effect size.

- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the journal of applied psychology: Little evidence of reform. *Educational and Psychological Measurement*, 61, 181–210.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational & Psychological Measurement*, 61, 213–218.
- Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan*, 67(1), 57–60.
- Thompson, B. (1994). The concept of statistical significance testing. *Practical Assessment, Research & Evaluation*, 4(5). Retrieved November 1, 2005, from <http://PAREonline.net/gettelevision.asp?v=4&n=5>
- Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, 62(2), 157–176.

SIMPLE MAIN EFFECT

Table 1 presents hypothetical data from a factorial design. Observations in each cell are averaged to produce a cell mean. Averaging over all levels of one factor produces means for the overall main effect of the other factor. For example, the six values observed with Drug 1 have a mean of 8.0 averaged over both dosage levels. Similarly, the six values from Drug 2 have a mean of 3.0. The difference between 8.0 and 3.0 is the overall main effect of the drug factor.

The two drugs could be compared separately for each dosage level. The difference between Drug 1 and Drug 2 in the 10 mg condition is 3.0 ($= 7.0 - 4.0$). This difference is an example of a simple main effect. That is, we can define a *simple main effect* as a difference in the value of a dependent variable found between two levels of one factor when another factor is held constant. The difference between the two drugs in the 20 mg condition is 7.0 ($= 9.0 - 2.0$).

Simple main effects can also be defined for the other factor. The difference between the two dosage levels for Drug 1 is 2.0 ($= 9.0 - 7.0$). Similarly, the difference between the two dosage levels for Drug 2 is 2.0 ($= 4.0 - 2.0$).

Simple main effects are often tested after finding a significant interaction in a factorial design. However, this practice can be misleading. For example,

Table 1 Hypothetical Data, Means, and Sum of Squares for a 2×2 Design

	B_1 10 mg	B_2 20 mg	Overall
A_1	8.0	10.0	
Drug 1	7.0	9.0	
	<u>6.0</u>	<u>8.0</u>	
Means	7.0	9.0	$\bar{X}_{A_1} = 8.0$
SS	2.0	2.0	$N_{A_1} = 6$
A_2	5.0	3.0	
Drug 2	4.0	2.0	
	<u>3.0</u>	<u>1.0</u>	
Means	4.0	2.0	$\bar{X}_{A_2} = 3.0$
SS	2.0	2.0	$N_{A_2} = 6$
Overall	$\bar{X}_{B_1} = 5.5$ $N_{B_1} = 6$	$\bar{X}_{B_2} = 5.5$ $N_{B_2} = 6$	$\bar{X}_T = 5.5$

Note: SS = sum of squares; $\bar{X}$ = mean.

Berrin-Wasserman, Winnick, and Borod reported a 2×2 design with a significant interaction and no significant simple main effects. A more complex problem is illustrated by Friedman, Putnam, Ritter, Hamberger, and Berman, who reported a 3×4 factorial design with a significant interaction in which testing simple main effects led to an incorrect interpretation of the reason for the significant interaction.

—Philip H. Ramsey

Further Reading

- Berrin-Wasserman, S., Winnick, W. A., & Borod, J. C. (2003). Effects of stimulus emotionality and sentence generation on memory for words in adults with unilateral brain damage. *Neuropsychology*, 17, 429–438.
- Friedman, D., Putnam, L., Ritter, W., Hamberger, M., & Berman, S. (1992). A developmental event-related potential study of picture matching in children, adolescents, and young adults: A replication and extension. *Psychophysiology*, 29, 593–610.

SIMPSON'S PARADOX

Simpson's paradox arises in the analysis of cross-classified categorical data when contradictory conclusions about the directions of associations

between variables are reached as a result of several groups' being aggregated into one. A well-known example of this phenomenon is the graduate school admissions data from the University of California, Berkeley. During a particular time, 8,442 men applied for graduate school, as did 4,321 women. Each department made its own admissions decisions. In the end, approximately 44% of the male applicants and 35% of the female applicants were admitted. This looks like evidence of sex bias in admissions, and the university administration wished to determine which departments exhibited the most serious bias. Here we explore only a subset of the data, namely the six largest departments. The data, as reported by Freedman and colleagues, are in Table 1.

Looking at each department individually, it is apparent that there is no systematic bias against women; on the contrary, in all six departments, the percentage of women accepted is higher than, or almost equal to, the percentage of men accepted. Indeed, for department A, 82% of female applicants were admitted, compared to only 62% of male applicants. The department that is least favorable to women, department E, shows a discrepancy of only 4%. Yet over all six departments together, 45% of the male applicants were admitted, and only 30% of the female applicants.

Closer examination of the data reveals the source of this seemingly paradoxical situation. Departments A and B were relatively liberal in their admissions policies, accepting two thirds or more of the applicants.

Table 1 UC Berkeley Graduate School Admissions Data

<i>Major</i>	<i>Men</i>		<i>Women</i>	
	<i>Number of Applicants</i>	<i>% Admitted</i>	<i>Number of Applicants</i>	<i>% Admitted</i>
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Source: Freedman, Pisani, Purves, & Adhikari (1991).

And most men applied to those two departments. On the other hand, departments C through F were much harder to get into, accepting about a third or less of prospective graduate students. And women applied mostly to those highly competitive departments. The overall effect is that a smaller percentage of women than men were accepted into the graduate programs at Berkeley because they were aiming for more restrictive courses of study. By aggregating over departments, university officials lost vital information.

This is the source of Simpson's paradox, namely, aggregating, or collapsing, over levels in cross-tabulated data when it is inappropriate to do so because of interactions among variables. The presence of interactions means that important and relevant information is lost when the table is collapsed. In the simple example of the graduate admissions data, the problem was easy to spot because the data were subjected to only a two-way classification. As the number of cross-classifications rises, inducing more complex and more numerous types of interactions, detection might not be so immediate. Care should be taken, therefore, when aggregating over levels of classification.

The phenomenon is called paradoxical because it allows researchers to make statements such as "The new treatment is more effective than the old treatment for men. It also works better than the old treatment for women. However, overall, the new treatment is not more effective than the old treatment." While apparently paradoxical, the result is a simple consequence of relationships between positive integers, as we will see in Table 2, a 2×2 table that looks at the proportions of males and females who improved with the old treatment versus the new.

Table 2 Mathematical Explanation of Simpson's Paradox

	<i>Improved With Old Treatment</i>	<i>Improved With New Treatment</i>
Men	a/b	c/d
Women	e/f	g/h
Total	(a+e)/(b+f)	(c+g)/(d+h)

In this setting, Simpson's paradox states that it is possible to find numbers such that $a/b < c/d$ and $e/f < g/h$ but $(a + e)/(b + f) > (c + g)/(d + h)$. Such values are not hard to find. For example, suppose the old treatment was tried on 200 men, of whom 30 improved, for a success rate of 15%, and on 400 women, of whom 300 improved, a success rate of 75%. On the other hand, the new treatment was tried on 200 men, of whom 40 improved, a success rate of 20%, and on 100 women, of whom 85 improved, a success rate of 85%. For both men and women, the new treatment was more efficacious than the old in terms of proportion who improved. But the old treatment was tried on a total of $200 + 400 = 600$ people, of whom $30 + 300 = 330$ improved, an overall improvement rate of 55%, whereas the new treatment was tried on only $200 + 100 = 300$ people, of whom $40 + 85 = 125$ improved, an overall rate of about 42%. Mathematically, these numbers fit the constraints described above, and similarly any other eight numbers that fit the constraints would result in a Simpson's paradox. Note that it is not the simple ratio that is relevant—substituting 4 for 40 and 20 for 200 for the men who received the new treatment would not meet the constraint, and hence there would be no paradoxical result.

More generally, the values that fit the constraints fall within two extreme conditions. On one end, we have situations in which slightly more units of one type are in a category with much lower probability; for example, see Table 3. Here, men are *much* less likely to improve in response to either treatment than are women, and the new treatment is given to slightly more men. As a result, a higher proportion of both men and women improve with the new treatment, but overall the old treatment is slightly more effective.

Table 3 One Scenario in Which Simpson's Paradox Holds

	<i>Improved With Old Treatment</i>	<i>Improved With New Treatment</i>
Men	1/45	5/55
Women	50/55	45/45
Total	51/100	50/100

At the other extreme are situations in which many more units of one type are in a category with slightly lower probability; for example, see Table 4. Here, both men and women have good rates of improvement, although the rate for the new treatment is slightly better. Men, with the lower probability of improvement, are overwhelmingly given the new treatment, whereas women, with a slightly better chance of improvement, are overwhelmingly given the old treatment. As a result, a higher proportion of both men and women improve with the new treatment, but overall the old treatment is slightly more effective.

Table 4 Another Scenario in Which Simpson's Paradox Holds

	<i>Improved With Old Treatment</i>	<i>Improved With New Treatment</i>
Men	4/5	90/95
Women	94/95	5/5
Total	98/100	95/100

In short, Simpson's paradox is the result of ignoring lurking, or background, variables that contain relevant information. The consequence of ignoring them is that contradictory conclusions may be reached regarding the direction of an association between variables. On an interesting historical note, a case may be made, according to S. E. Fienberg, that the paradox should in fact be named after Yule (and is indeed sometimes called the Yule-Simpson paradox), who apparently discussed it as early as 1903.

—Nicole Lazar

Further Reading

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Fienberg, S. E. (1994). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York: Norton.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238–241.

Simpson's paradox: http://exploringdata.cqu.edu.au/sim_par.htm

Simpson's paradox entry from the *Stanford Encyclopedia of Philosophy*: <http://plato.stanford.edu/entries/paradox-simpson>

SIMPSON'S RULE

In order to estimate the integral of a function of one real variable, f , over a finite interval $[a, b]$, let $h = (b - a)/n$ for a positive even integer $n = 2m$. Then the (composite) Simpson's rule formula is

$$\int_a^b f(x) dx \approx (h/3) (f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + 2f(a+4h) + \dots + 2f(b-2h) + 4f(b-h) + f(b)) \quad (1)$$

with error, assuming f'''' exists on (a, b) , given by

$$- (b - a) h^4 f''''(c)/180 \quad (2)$$

for some c in (a, b) . The key features of the error term are the dependence on h^4 and the $f''''(c)$ term. These imply that doubling the number of subintervals (so that h is halved) yields approximately 1/16 the previous error (if the overall interval $[a, b]$ is sufficiently small so that c does not vary too greatly), and the integral will be exact for polynomials of degree 3 or less, since f'''' vanishes identically for such functions.

(Note that a method known as *Richardson extrapolation* can be used to improve Simpson's rule integral estimates even further if two different h values are used.)

As an example, the probability that a standard normal random variable has value between $z = .5$ and $z = 1$ is given by the integral of $f(z) = \exp(-z^2/2)/\sqrt{2\pi}$ from $z = .5$ to $z = 1$. Choosing $n = 10$, Simpson's rule estimates it to be $(.05/3)(f(.5) + 4f(.55) + 2f(.6) + 4f(.65) + \dots + 4f(.95) + f(1))$ or 0.14988228478. . . . Now $f''''(z)$ is $(z^4 - 6z^2 + 3)f(z)$,

and with c between .5 and 1, the error term is bounded from above by $.6(.56)(.05^4)/180$, that is, by 10^{-8} . (The exact error is, in fact, 1.3×10^{-11} .)

The logic underlying Simpson's rule considers f on $m = n/2$ equally spaced contiguous nonoverlapping subintervals, and on each subinterval of width $2h$, it estimates the integral by using a quadratic interpolation. From the partition $\{a = x_0, x_1, \dots, b = x_{2m}\}$, where $x_{i+1} - x_i = h$, the integral of f on $[x_{2j}, x_{2j+2}]$ for $j = 0$ to $m - 1$ is approximated by the integral of the quadratic function q_j satisfying $q_j(x_{2j}) = f(x_{2j})$, $q_j(x_{2j+1}) = f(x_{2j+1})$, and $q_j(x_{2j+2}) = f(x_{2j+2})$. By elementary calculus, the integral of q_j on $[x_{2j}, x_{2j+2}]$ is exactly $(f(x_{2j}) + 4f(x_{2j+1}) + f(x_{2j+2}))(h/3)$; it is more difficult to deduce that the error incurred on the subinterval is $-f''''(c)h^5/90$ for some c in (x_{2j}, x_{2j+2}) . Summing up estimates from all the subintervals yields Equation 1; collecting the errors and using the intermediate value property for derivatives gives the error term (Equation 2).

In its modern formulation, Simpson's rule is one of an infinite family of what are known as (closed-form) *Newton-Cotes integration methods*. (The trapezoid rule is another, less accurate, closed-form Newton-Cotes method; the midpoint rule is an open-form Newton-Cotes method.) Each estimates integrals by computing the exact integral of an appropriate (piecewise) interpolating polynomial. Using any degree of interpolant, one can quickly deduce any Newton-Cotes weights, such as the $h/3$, $4h/3$, $h/3$ for Simpson's rule, and so deduce ever more accurate integration schemes, as follows. Requiring that the integration method be exact for polynomials of degree d or lower and that the function be evaluated at equally spaced intervals leads quickly to linear conditions for the weights. For instance, in order to achieve exact integration results for polynomial integrands of degree through $d = 2$, consider the monomials $f(x) = 1$, x and x^2 on the interval $[0, 2h]$: Equating the integral with $w_0 f(0) + w_1 f(h) + w_2 f(2h)$ for these functions yields the system $2h = w_0 + w_1 + w_2$, $2h^2 = h w_1 + 2h w_2$, $8h^3/3 = h^2 w_1 + 4h^2 w_2$, whose solution is $w_0 = h/3 = w_2$, $w_1 = 4h/3$. To determine the error term, assuming it has the form $K f^{(n)}(c) (b - a)h^k$ for some K , n , and k , apply the estimation to, for instance, $f(x) = x^4$ on $[0, 2h]$; since $f''''(c) = 24$ for any c , and the error

incurred is $-4h^5/15$, the error term must be $-h^4(b-a)f''''(c)/180$. Similarly, considering the related example of exact results for monomial integrands $f(x) = 1, x, x^2$ and x^3 on $[0, 3h]$, one obtains the weights $w_0 = 3h/8$, $w_1 = 9h/8$, $w_2 = 9h/8$, $w_3 = 3h/8$ (known as Simpson's "3/8" rule). Using $f(x) = x^4$, the overall error term is found to be $-h^4(b-a)f''''(c)/80$. (Note that the 3/8 rule, while based on a higher-degree interpolant than Simpson's rule, has a larger coefficient in its error term and ordinarily is not preferred to Simpson's rule. Its advantage is that it allows use of three subintervals. Thus, if data are supplied for an odd number of subintervals, the 3/8 rule can be used for the first three subintervals, and Simpson's rule can be applied to the others.)

Simpson's rule arguably appeared by 1890 B.C. in what is known as the Russian papyrus and was used for measuring volumes of prisms. It was known by Johannes Kepler (1571–1630) as an aid in estimating volumes of barrels. James Gregory (1638–1675) published an algebraic formulation in 1668. As a method developed by Isaac Newton, it appeared in Thomas Simpson's calculus text *A New Treatise on Fluxions* in 1737. Variants of Simpson's rule that have most weights equal to 1 have been found, and Simpson's rule can be used to sum series as well as to estimate integrals.

—Richard M. Kreminski

Further Reading

Burden, R. L., & Faires, J. D. (2004). *Numerical analysis* (8th ed.). Pacific Grove, CA: Brooks/Cole.

MacTutor History of Mathematics archive: <http://turnbull.mcs.st-and.ac.uk/~history/>

SIMULATED ANNEALING

Simulated annealing (SA) is an optimization technique that mimics the physical process of thermal cooling. In physics, annealing is the process in which one first heats up a solid to a high enough temperature to liquefy it and then cools the liquid slowly enough

so that the particles arrange themselves in a lattice arrangement having the lowest energy.

Statistical physics describes this cooling process in probabilistic terms by describing the probability that a particle will be in a state with a given energy E and for a fixed temperature T . This probability is given as

$$Pr(\text{Energy} = E) = \frac{1}{Z(T)} \exp\left(-\frac{E}{k_B T}\right),$$

where k_B is the *Boltzmann constant* and $Z(T)$ is a normalizing constant, known as the *partition function*. This distribution is known as the *Boltzmann distribution*.

Looking at this probability, it is clear that as T tends to zero, the distribution concentrates in the minimal energy states. In the limit as $T \rightarrow 0$, we see that the only possible states are the minimal energy states.

The first computer algorithm for simulating the process of thermal cooling was developed by Metropolis and colleagues in 1953. However, not until 1982 was the link between abstract combinatorial optimization and the minimization of energy in thermal cooling explicitly made and exploited.

The idea is to exploit the temperature dependence of the Boltzmann distribution to try to force the current state of our simulation (optimization algorithm) toward one of the globally optimal (minimal energy) states. The key lies in carefully lowering the temperature.

The Metropolis Algorithm

In 1953, N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller published the paper "Equations of State Calculations by Fast Computing Machines," thereby introducing to the world the highly useful Metropolis simulation algorithm. This algorithm has been cited as among the top 10 algorithms having the greatest influence on the development of science and engineering.

This algorithm will yield the Boltzmann distribution for a given system via simulation.

The basic idea, like many great ideas, is rather simple. Given the energy function E on the state space

Ω , we split the chain into a *proposal* phase and an *acceptance* phase. For the proposal phase, the current state x of the system is subjected to a small perturbation, which generates y , a proposed new state. This proposal process needs to be symmetric (that is, it is just as likely to propose x starting from y as it is to propose y starting from x) and irreducible (it is possible to transition from any state to any other state using only the proposal process). Let $E(x)$ and $E(y)$ denote the energies associated with states x and y . If $E(y) < E(x)$, then we accept the proposed change, and y becomes the new current state. If $E(y) \geq E(x)$, we accept y with probability

$$\exp\left(-\frac{E(x) - E(y)}{k_B T}\right),$$

else we reject the new state y and stay with the original state x .

The Markov chain X_n generated by the Metropolis algorithm will produce many repeated states. In computing using this chain, it is essential to keep these repeated states because this is one way the algorithm corrects the bias from the proposal process, thus allowing it to converge to the Boltzmann distribution.

The separation of the proposal and acceptance phases makes designing a chain much simpler because the proposal process can be virtually any process. The choice of proposal process is often guided by the structure of the state space. The conditions on the proposal process are rarely a problem in practice. For Ω finite, the proposal process is often defined by first giving a local (symmetric) neighborhood structure and then choosing y uniformly from the neighbors of the current x . As long as the neighborhood structure is reasonable, this process will usually work.

We are free to let the structure of the problem guide the neighborhood system (or proposal process) and use the acceptance phase to correct the proposal process in such a way as to ensure the Boltzmann distribution. Heuristically, the neighborhood system defines a type of “topology” on the state space Ω , and along with the energy function, this gives a type of “energy landscape.”

The convergence rate of the Metropolis algorithm is strongly influenced by the proposal process. Thus, any problem-specific information is very useful in designing an efficient Metropolis scheme.

Notice that the Metropolis algorithm does not use the partition function $Z(T)$. This is very important as it is usually impossible to compute this normalizing constant because the state space is usually prohibitively large. The Metropolis algorithm uses only energy differences, so really it uses information about only relative probabilities (under the Boltzmann distribution) and not absolute probabilities.

SA as Time-Varying Metropolis

For a fixed temperature $T > 0$, the Metropolis algorithm gives us a way to simulate a distribution in which better (lower-energy) states are more likely to occur than worse (higher-energy) states. As mentioned above, clearly lowering the temperature will concentrate the Boltzmann distribution more in the low energy states, so the idea behind SA is to lower the temperature in a careful way.

The simplest way to lower the energy would be to have a sequence of temperatures $T_1 > T_2 > T_3 > \dots$ that approaches zero. Start with some state X_0 and run the Metropolis algorithm at temperature T_1 until it has converged to the Boltzmann distribution for temperature T_0 , say some number N_1 of iterations, so that we are in state X_{N_1} . Then using X_{N_1} as the initial state, run the Metropolis algorithm at temperature T_2 for N_2 iterations until it converges. Continuing, we piece together runs of the Metropolis algorithm at each different (but fixed) temperature T_i to get a sequence of states that should converge to a globally minimal state.

Simulated annealing does something like this, but with a temperature that decreases every iteration according to some *cooling schedule*. Thus, an SA process never really reaches an equilibrium distribution because the equilibrium Boltzmann distribution is changing as a function of the current temperature. The idea is that if you decrease the temperature slowly enough, then you track the current Boltzmann distribution closely enough that the overall process will converge.

In general, the *relaxation time* increases as the temperature decreases. As we lower the temperature, the algorithm requires a larger number of iterations to approach the equilibrium distribution. Thus, the convergence of the algorithm slows down considerably for lower temperatures.

If the temperature is lowered too quickly, the system can be “frozen” into a metastable state. In terms of the objective function, this means that we are trapped into a local but not global minimum value of the function. As the algorithm progresses, the current point moves around the state space by taking small steps (the perturbations). If the proposed step is “downhill,” then the step is always taken. If the proposed step is “uphill,” then we sometimes take it. Initially, at high temperatures, many uphill steps are taken. This allows the system to escape from the energy well of a local minimum. As the temperature decreases, the probability of taking an uphill step decreases, so it is more difficult to escape from the basin of a local minimum. If the temperature is decreased too fast, it is possible to get stuck in a local minimum basin, and then the state will converge to the local minimum.

Convergence Results

One of the very nice things about SA is that there are theoretical results that guarantee convergence to a minimal energy state, albeit in the infinite time limit. The 1988 result of Hajek is one such result. Letting T_n denote the temperature at iteration n , a *logarithmic cooling schedule* is of the form

$$T_n = \frac{c}{\log n + 1}.$$

If $c > 0$ is large enough, then Hajek’s result says that using the logarithmic cooling schedule results in $\Pr(X_n \text{ is an optimal state}) \rightarrow 0$ as $n \rightarrow \infty$. How large is large enough? Well, if we think of the objective function as defining an energy landscape, then c should be at least as large as the largest “energy barrier” on the surface, that is, the depth of the largest nonoptimal local optimal “valley.”

There are other results that guarantee convergence or give estimates on the convergence rate. However, in

the general situation, all the convergence results essentially require a logarithmic cooling schedule.

This logarithmic cooling schedule can be understood as saying that, in general, the relaxation times grow geometrically as the temperature decreases. Thus, you need something like twice as long at a lower temperature in order to approach the equilibrium.

Example

It is easy to set up an example in which the expected time to convergence is infinite. The following example was given by Shonkwiler and Van Vleck. Let $\Omega = \{0, 1, 2\}$ with function values $f(0) = -1$, $f(1) = 1$, and $f(2) = 0$ and proposal matrix

$$\begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

The Metropolis transition matrix then is

$$P(T) = \begin{pmatrix} 1 - \frac{1}{2}e^{-2/T} & \frac{1}{2}e^{-2/T} & 0 \\ 1/2 & 0 & 1/2 \\ 0 & \frac{1}{2}e^{-1/T} & 1 - \frac{1}{2}e^{-1/T} \end{pmatrix}.$$

Using the logarithmic cooling schedule $T_n = 1/\log(n+1)$, we see that this translates to

$$P(n) = \begin{pmatrix} 1 - \frac{1}{2(n+1)^2} & \frac{1}{2(n+1)^2} & 0 \\ 1/2 & 0 & 1/2 \\ 0 & \frac{1}{2(n+1)} & 1 - \frac{1}{2(n+1)} \end{pmatrix}.$$

Letting h_k be the probability that the first time we are in State 0 is at time k , it can be shown that $h_k \geq \frac{1}{4k(k-1)}$ for $k \geq 2$, and thus the expected hitting time starting from State 2 is at least as large as

$$\sum_{k=2}^{\infty} kh_k \geq \frac{1}{4} \sum_{k=2}^{\infty} \frac{1}{k-1} = \infty,$$

so the expected hitting time for SA is infinite.

Cooling Schedules

While the logarithmic cooling schedule is sufficiently slow to guarantee convergence of SA, it is clearly extremely slow. In fact, using a logarithmic cooling schedule is in effect often the same as performing an exhaustive search (even worse than exhaustive search, in the above simple three-state example). Furthermore, in practice one obviously cannot run an algorithm until the infinite time limit. Thus many other cooling schedules are used. A particularly simple one is geometric cooling, of the form $T_n = T_0 \alpha^n$, where $0 < \alpha < 1$.

It might also be desirable to have the temperature decrease adaptively. To do this, one can accumulate statistics about the process (distribution of energy values encountered, number of proposed steps accepted/rejected, etc.) and use this information to derive a model of the energy surface for the problem.

A particularly nice approach to this is one derived from finite-time statistical mechanics. Since the logarithmic cooling schedule and infinite time runs are not practical, we use finite-time and nonequilibrium statistical mechanics to derive “optimal” cooling schedules, such as the *constant thermodynamic speed* cooling schedule. However, these cooling schedules require a fair bit of information about the energy surface. Depending on the optimization problem, it may be better to run many instances of a simpler cooling schedule (or longer runs) than one run of an adaptive cooling schedule.

Application of SA to the Traveling Salesman Problem

The traveling salesman problem (TSP) is a classical combinatorial optimization problem that is NP-hard. The data for the TSP are the geographical locations of N cities and a notion of the “cost” of traveling between any two of these cities. This cost is typically associated with the distance between the cities but does not have to be. Given these N cities, we are to find the tour (list of which cities to visit in order) that minimizes the total cost. For simplicity, we assume that we start and end at the first city (clearly any city on a tour can be thought of as the “first” city on the tour). The N -city TSP has $(N-1)!/2$

distinct tours, so the state space is incredibly huge, even for moderate N .

The first step in designing an SA algorithm for this problem is to decide on the proposal process. This is a crucial step because the convergence rate of the Metropolis algorithm depends very heavily on the design of the proposal process. For our example, we use the 2-opt proposal process for the TSP. To explain this, given the tour

$$c_1 c_2 c_3 \dots c_{a-1} \underbrace{c_a c_{a+1} \dots c_{b-1} c_b}_{\text{segment}} c_{b+1} \dots c_N$$

we choose two positions, $a, b \in \{1, 2, \dots, N\}$, and “flip” the part of the tour between these positions to get the proposed tour

$$c_1 c_2 c_3 \dots c_{a-1} \underbrace{c_b c_{b-1} \dots c_{a+1} c_a}_{\text{segment}} c_{b+1} \dots c_N.$$

As a more concrete example, the tour 136425 could be perturbed into the tour 132465 by choosing locations 3 and 5.

This 2-opt proposal process has the feature of breaking two “links” of the tour and replacing them with two other (new) links.

Clearly, the energy for the TSP will be the total cost. Since we are trying to minimize the total cost, this will work as an energy (SA naturally minimizes; a maximization problem needs to be converted to a minimization problem).

Figure 1 shows two typical runs of SA on the TSP. In both figures, the horizontal axis represents iterations and the vertical axis represents energy. The bottom curve is the best energy seen so far, and the top one is the energy of the current state. These figures illustrate that when the temperature is high, the energy of the current state is quite variable, but as the temperature decreases, the energy of the current state tends to settle down and “freeze.” All the runs used a fixed set of 100 cities distributed somewhat randomly in the plane.

Figure 2 shows the results of 10 runs of SA on the TSP. These plots show only the trace of the best so far.

Premature Convergence

An individual run of SA will typically have a short period of rapid decrease followed by long, flat periods

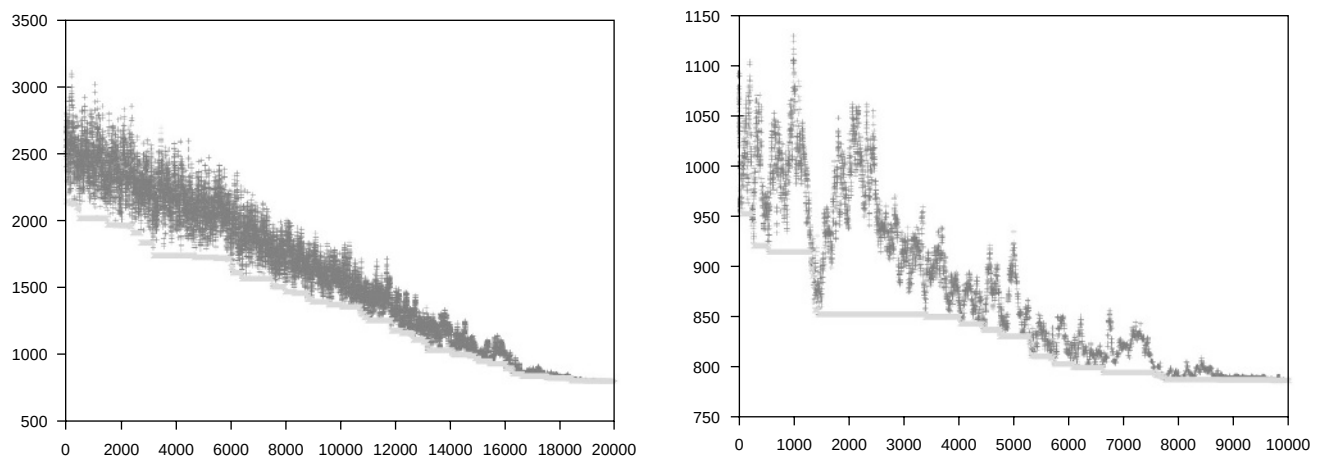


Figure 1 Two Sample Paths From SA on the TSP

in which the best energy stays constant. These flat periods can eat up a majority of the computational effort, especially for very long runs.

There are several different strategies for dealing with this problem. One simple problem is to randomly restart the SA process whenever such a situation occurs. One method that works well is to restart whenever the current state has not changed in some set number of iterations, the idea being that this indicates that a large majority of the “neighbors” of the current state are worse than the current state. Another method, which is also effective, is to run several versions of the algorithm in parallel and do some sort of information sharing.

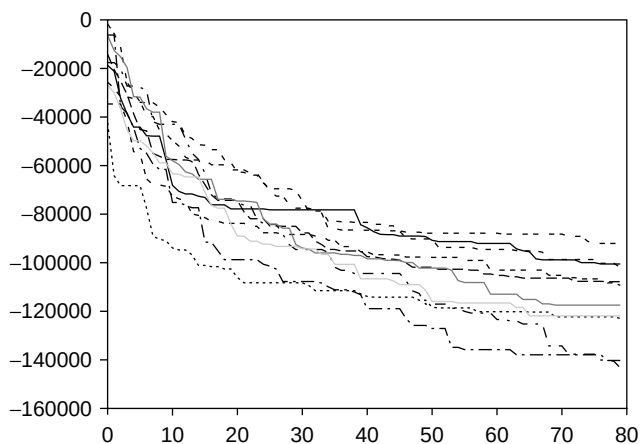


Figure 2 Ten Runs of SA on the TSP (Best So Far)

SA is a natural candidate for parallelization. In fact, it allows the simplest type of parallelization—multiple independent runs of the identical algorithm, but from randomly chosen starting conditions. Many other types of parallelization involving some amount of information sharing are also possible and effective.

—Franklin Mendivil

See also EM Algorithm; Markov Chain Monte Carlo Methods

Further Reading

- Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. New York: Springer.
- Catoni, O. (1991). Sharp large deviations estimates for simulated annealing algorithms. *Annales de l'institut Henri Poincaré, Probabilités et Statistiques*, 27(3), 291–383.
- Givens, G., & Hoeting, J. (2005). *Computational statistics*. New York: Wiley-Interscience.
- Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Norwell, MA: Kluwer.
- Nourani, Y., & Andresen, B. (1998). A comparison of simulated annealing cooling strategies. *Journal of Physics A: Mathematical & General*, 31, 8373–8385.
- Nulton, J. D., & Salamon, P. (1988). Statistical mechanics of combinatorial optimization. *Physical Review A*, 37(4), 1351–1356.

Salamon, P., Nulton, J., Harland, J. R., Pedersen, J., Ruppeiner, G., & Liao, L. (1988). Simulated annealing with constant thermodynamic speed. *Computer Physics Communications*, 43, 423–438.

Seneta, E. (1981). *Non-negative matrices and Markov chains*. New York: Springer.

SIMULATION EXPERIMENTS

Simulation experiments are used to mimic a system of interest and are typically, although not necessarily, performed on a computer. A simulation experiment differs from a “simulator” (e.g., a flight simulator), in which an operator is immersed in a virtual environment. In a simulation experiment, the modeler gleans the appropriate information about the system of interest (which may include data gathering), then develops equations and algorithms to simulate the system. These equations and algorithms are then converted to a computational model, which is typically implemented on a digital computer for analysis.

This brief entry provides a description of the mathematical and  computational techniques for modeling, simulating, and analyzing the performance of stochastic systems. By definition, the nature of simulation is that one does not actually experiment with or modify an actual system. Instead, one develops and then works with a mathematical model of the actual system. This model allows the modeler to respond to “what if” questions associated with the model, and these questions, if the model has been developed properly, translate to the associated questions concerning the system.

Model Characterization

A simulation model is typically both stochastic and dynamic, as described in the paragraphs that follow.

A system model is either deterministic or stochastic. A deterministic system model has no stochastic (random) components. For example, provided the conveyor belt and machine never fail, a model of a constant velocity conveyor belt feeding parts to a machine with a constant service time is deterministic.

At some level of detail, however, all systems have some stochastic components: Machines fail, people respond in a random fashion, and so on. One attractive feature of simulation modeling is that stochastic components can be accommodated, usually without a dramatic increase in the complexity of the system model at the computational level.

A system model is static or dynamic. A static system model is one in which time is not a significant variable. Many static models can be analyzed by Monte Carlo simulation, which performs a random experiment repeatedly to estimate, for example, a probability associated with a random event. The passage of time always plays a significant role in dynamic models.

A dynamic system model is continuous or discrete. Most of the traditional dynamic systems studied in classical mechanics have state variables that evolve continuously. A particle moving in a  gravitational field, an oscillating pendulum, and a block sliding down an inclined plane are examples. In each of these cases, the motion is characterized by one or more differential equations that model the continuous time evolution of the system. In contrast, models involving queuing, machine repair, inventory systems, and so forth, are discrete because the state of the system is a piecewise-constant function of time. For example, the number of jobs in a queuing system is a natural state variable that changes value only at those discrete times when a job arrives (to be served) or departs (after being served).

A discrete-event simulation model is defined by three attributes: (a) The model is stochastic—at least some of the system state variables are random; (b) the model is dynamic—the time evolution of the system state variables is significant; and (c) the model has discrete event times—significant changes in the system state variables are associated with events that occur at discrete time instances only.

We focus the balance of this entry on discrete-event simulation because of its utility and popularity as a tool used by practitioners and researchers from several application areas (e.g., operations research, computer science, biomathematics, and statistics).

Simulation Model Development

We present a series of steps that describe, at a high level, how to develop a discrete-event simulation model and then conduct a discrete-event simulation study.

A typical discrete-event simulation model will be developed consistent with the following six steps. Steps 2 through 6 are typically iterated, perhaps many times, until a valid computational model (a computer program) has been developed.

1. Set goals. Once a target system of interest has been identified, determine the goals and objectives of the analysis. These goals and  objectives are often phrased as simple Boolean decisions (e.g., should an additional server be added to a queuing network service node?) or numeric decisions (e.g., how many parallel servers are necessary to provide satisfactory performance in a multiple-server queuing system?). Without specific goals and objectives, the remaining steps lack meaning.

2. Build a conceptual model of the system based on Step 1. What are the state variables, how are they interrelated, and to what extent are they dynamic? How comprehensive should the model be? Which state variables are important? Which have such a negligible effect that they can be ignored? In effect, establish a boundary for what is included in the mathematical model of the system and what is considered part of the external environment.

3. Convert the conceptual model into a specification model. This step typically involves collecting and statistically analyzing data to provide the stochastic “input models” that drive the simulation. In the absence of such data, the input models must be constructed in an ad hoc manner using stochastic models believed to be representative.

4. Turn the specification model into a computational model (a computer program). At this point, a fundamental choice must be made—to use a general-purpose programming language or a special-purpose simulation language.

5. Verify the model. As with all computer programs, the computational model should be consistent with the specification model.

6. Validate the model. Is the computational model consistent with the system being analyzed? Because the purpose of simulation is insight, some would argue that the act of developing the discrete-event simulation model—Steps 2, 3, and 4—is frequently as important as the tangible product. One popular non-statistical technique for model validation is to place actual system output alongside similarly formatted output from the computational model. This output is then examined by an expert familiar with the system. Model validation is indicated if the expert is not able to determine which is the model output and which is the real thing. Interactive computer graphics (animation) can be very valuable during the verification and validation steps.

The actual development of a complex discrete-event simulation model may not be as sequential as the previous steps suggest, particularly if the development is a team activity (in which case some steps will probably be worked on in parallel). The different characteristics of each step should always be kept clearly in mind, avoiding, for example, the natural temptation to merge Steps 5 and 6.

After the successful creation of a computational model (computer program) by means of Steps 1–6, the analysis of that computational model involves the following steps:

7. Design the simulation experiments. If there are a significant number of system parameters, each with several possible values of interest, then the combinatoric possibilities to be studied need to be managed.

8. Make production runs. The simulation runs should be made systematically, recording the value of all initial conditions and input parameters, along with the associated statistical output. Point and confidence-interval estimators of measures of performance of interest are typical discrete-event simulation outputs.

9. Analyze the simulation results. The analysis of the simulation output is statistical in nature because discrete-event simulation models have stochastic (random) components. Common statistical analysis tools (means, standard deviations, percentiles, histograms, confidence intervals, correlations, etc.) are used to draw statistical inferences concerning the model.

10. Make decisions. It is to be hoped that the results of Step 9 will lead to decisions that result in actions. If so, the extent to which the computational model correctly predicted the outcome of these actions is always of great interest, particularly if the model is to be further refined in the future.

11. Document the results. This documentation should include assumptions made concerning the model, details associated with translating the model into computer code, and a summary of insights gained concerning the system. If decisions are made and associated action taken, then the results of that action on the system should be documented.

An important benefit of developing and using a discrete-event simulation model is that valuable insight concerning the system is acquired. As conceptual models are formulated, computational models developed, and simulation output data analyzed, subtle system features and component interactions are often discovered that would not have been noticed otherwise. The systematic application of the 11 steps outlined above can result in better actions taken as a result of insight gained by an increased understanding of how the system operates.

Simulation Programming Languages

There is a continuing debate in the discrete-event simulation community over whether to use a general-purpose programming language or a (special-purpose) simulation programming language. For example, two standard discrete-event simulation textbooks provide the following conflicting advice. Bratley, Fox, and Schrage state, "For any important large-scale real application we would write the programs in a standard general-purpose

language, and avoid all the simulation languages we know." In contrast, Law and Kelton state, "We believe, in general, that a modeler would be prudent to give serious consideration to the use of a simulation package." General-purpose languages are more flexible and familiar; simulation languages allow modelers to build computational models quickly.

Simulation languages have built-in features that provide many of the tools needed to write a discrete-event simulation program. Because of this, simulation languages support rapid prototyping and have the potential of decreasing programming time significantly. Moreover, animation is a particularly important feature now built into most of these simulation languages. This is important because animation can increase the acceptance of discrete-event simulation as a legitimate problem-solving technique. By the use of animation, dynamic graphical images can be created that enhance verification, validation, and the development of insight. (The features and costs of these languages are surveyed by Swain.)

—Lawrence Leemis

Further Reading

- Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-event system simulation* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Bratley, P., Fox, B. L., & Schrage, L. E. (1987). *A guide to simulation* (2nd ed., p. 219). New York: Springer-Verlag.
- Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed., p. 204). New York: McGraw-Hill.
- Leemis, L. M., & Park, S. K. (2006). *Discrete-event simulation: A first course*. Englewood Cliffs, NJ: Prentice Hall.
- Swain, J. J. (2005). Seventh Biennial Survey of Discrete-Event Software Tools. *OR/MS Today*, 32.

SINGLE-SUBJECT DESIGNS

Single-subject designs involve the intensive study of individual subjects under various conditions, in various environments, or some combination thereof. The subject of interest is often a single organism,

such as a client in a clinical setting, but could also be a group of individuals acting as a single unit, such as a department within an organization or a school within a district. This focus on the individual is reflected in the alternative labels for single-subject designs, including *single-case designs*, *N of 1 designs*, and *within-subject designs*. Although single-subject designs focus on the individual, studies using these designs typically include several participants.

Single-subject designs can be used in correlational or experimental studies, but we will focus on experimental designs. These designs share some basic features with group (or *between-subject*) experimental designs: An independent variable (IV) is manipulated, a dependent variable (DV) is measured, and nuisance variables (NVs) are controlled. Unlike participants in group designs, participants in single-subject designs receive *all levels* of the IV, while the DV is measured repeatedly under those levels. The values of the DV under each level of the IV (e.g., a control condition and one or more treatment conditions, or *interventions*) are compared, and causal inferences regarding the effects of the intervention(s) are then made. In these designs, each participant serves as its own control.

Single-subject designs provide useful and flexible methodological options when (a) one cannot obtain appropriate numbers of participants for group designs (e.g., the population of interest is relatively small, such as persons with a rare medical condition), (b) one seeks to assess treatment effectiveness in particular individuals (e.g., the progress of specific clients is of interest), or (c) one cannot ethically use group designs (e.g., withholding treatment from individuals in a control group is not an option). Practical reasons may also necessitate the use of single-subject designs.

The specific details of single-subject designs vary, but most such designs share two common features. First, the DV is measured repeatedly across time; how often it is measured depends on many factors, including available resources and the nature of the study. Repeated measurement permits the researcher to track changes in the DV and, if

necessary, to modify treatment conditions if the DV does not change in the desired direction. Second, the level of the DV in the absence of treatment is established during a *baseline* (or *A*) *phase*, which typically occurs at the beginning of the study, before treatment is administered, although there are some exceptions.

The baseline phase serves three important functions: (a) It provides information regarding the pre-treatment level of the DV; (b) it provides information on the predicted level of the DV if treatment is not applied; and (c) it provides control data with which treatment data will be compared, with differences in the DV between baseline phase and the treatment (or *B*) phase providing evidence for the effects of the intervention. Figure 1 presents fictitious data to illustrate these functions. The DV is the number of accidents per week at a job site, and the intervention is a program that provides workers with incentives for every day without an accident. Panel A shows that the baseline level of the DV hovers around 15 accidents per week. From these data, we can predict that the number of accidents would probably remain the same if the intervention were not introduced. Panel B depicts a decrease in the number of accidents on introduction of the intervention. The change in the DV from the baseline phase to the treatment phase provides evidence, albeit limited, of the intervention's effectiveness.

To aid the interpretation of an intervention's effectiveness, baseline data should show stability. *Stability* means the data show neither a systematic upward or downward trend (i.e., the data have a slope near zero) nor excessive variability (i.e., the data hover around a consistent value, showing only minor fluctuations from one measurement occasion to the next). Stable baseline data provide a relatively clear benchmark against which to assess the effects of the intervention and information about the steady-state level of the DV. There are no simple rules for deciding when baselines are sufficiently stable, for such a decision depends on several factors, including the nature of the DV, the subject matter, and practical considerations. Notice that Panels A and B of Figure 1 depict stable data. Panels C and D of

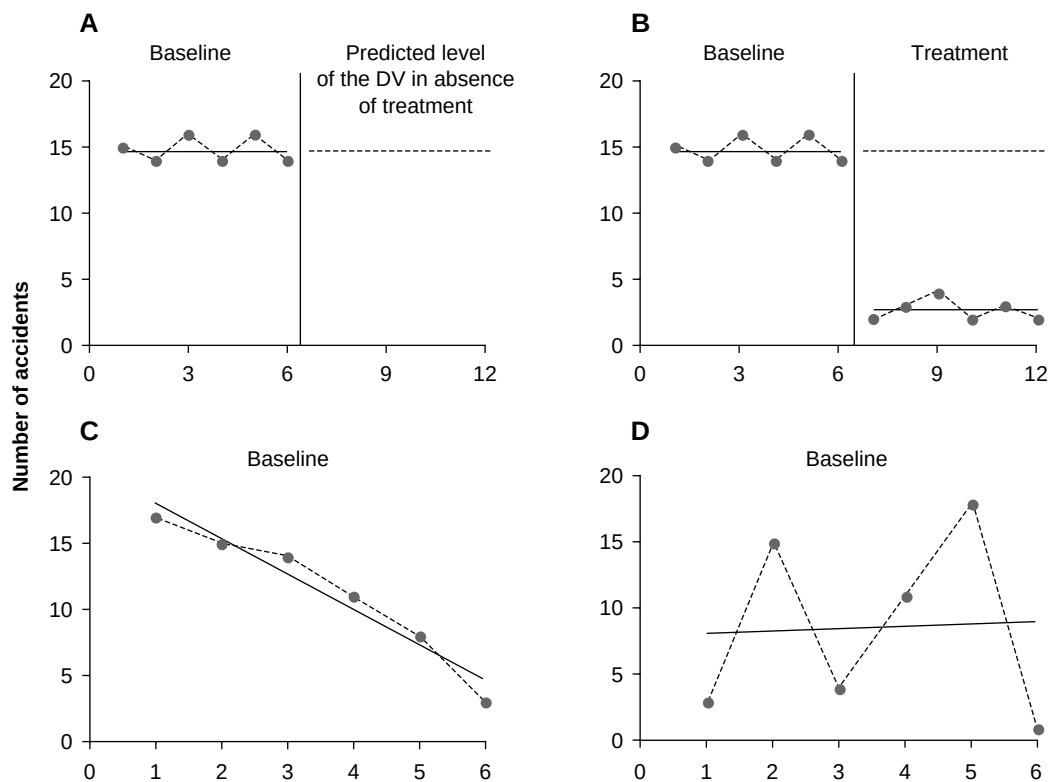


Figure 1 Data Illustrating Functions of Baseline Phase

Note: Panel A depicts a two-phase withdrawal design with baseline data and the level of the dependent variable predicted from the baseline data. It is assumed that in the absence of an intervention, the level of the dependent variable will remain similar to that seen in the baseline phase. Panel B depicts the same design with the addition of data demonstrating an effective treatment. The dashed line indicates the predicted level of the dependent variable in the absence of a treatment effect. The difference in the predicted level and the observed level of the dependent variable represents the size of the treatment effect (in this case, approximately 12 points). Panel C depicts a downward trend in the baseline data. Panel D depicts highly variable data.

Figure 1 depict baseline data that are unstable and extremely variable, respectively.

Withdrawal Designs

The example in Figure 1 represents a quasi-experimental single-subject design that consists of two phases, a baseline (A) phase followed by a treatment (B) phase. Such A-B designs are vulnerable to several threats to internal validity. It is possible that some NV, and not the intervention, caused the observed changes in the DV during the B phase. For example, these changes could have resulted from the coincidental occurrence of some external event or from changes in the participants themselves. To increase confidence that the

intervention was responsible for changes in the DV, the researcher could withdraw the intervention and return to baseline conditions, yielding an A-B-A design. Such designs are often termed *withdrawal designs* because the current condition is withdrawn and replaced with another condition.

Withdrawal designs help one rule out the possible effects of NVs on the DV by demonstrating systematic changes in the DV following introduction and withdrawal of the intervention. If the level of the DV in the second baseline phase returns to a level similar to that seen in the first baseline, one can be more confident that the intervention alone caused the observed change during the treatment phase, for it is unlikely that the actions of an NV happened to coincide with

both changes in phases. When changing phases, one should change only one condition at a time. Changing multiple conditions simultaneously makes the determination of the relative influence of each condition on the DV difficult, if not impossible.

One problem with A-B-A designs is that they leave participants in an untreated baseline phase. Thus, one may wish to reinstate the treatment after the second baseline phase, yielding an A-B-A-B design. In addition to leaving participants in a treatment phase, A-B-A-B designs can provide further evidence for treatment effects by demonstrating another change in the DV following the intervention's reintroduction. Two limitations of withdrawal designs are that (a) they cannot be used to study interventions with irreversible effects (e.g., skills training programs, certain medical procedures), and (b) they may require the repeated withdrawal of an effective treatment, which may be impractical, unethical, or impossible. In such situations, other single-subject designs can be used. One such design is the multiple-baseline design.

Multiple-Baseline Designs

As their name implies, multiple-baseline designs assess the effects of the intervention across separate baselines. These baselines can involve (a) different DVs in the same participant in one setting, (b) the same DV in the same participant in different situations, or (c) the same DV in different participants under similar or different conditions. In all three variations, measurements are made simultaneously on all baselines. When the DV on one baseline shows stability, the intervention is introduced (i.e., the treatment phase begins) while measurements continue on the other baselines. If the intervention produces changes only on the DV in the treatment phase and not on DVs in other baseline phases, we can feel more confident that the intervention, and not some other variable, was responsible for the observed change. If this is the case, and if the data are stable in the next baseline, then the intervention is introduced on the next baseline. In this way, the introduction of the intervention is staggered across baselines, and the multiple-baseline design thus comprises a collection of A-B designs. Figure 2

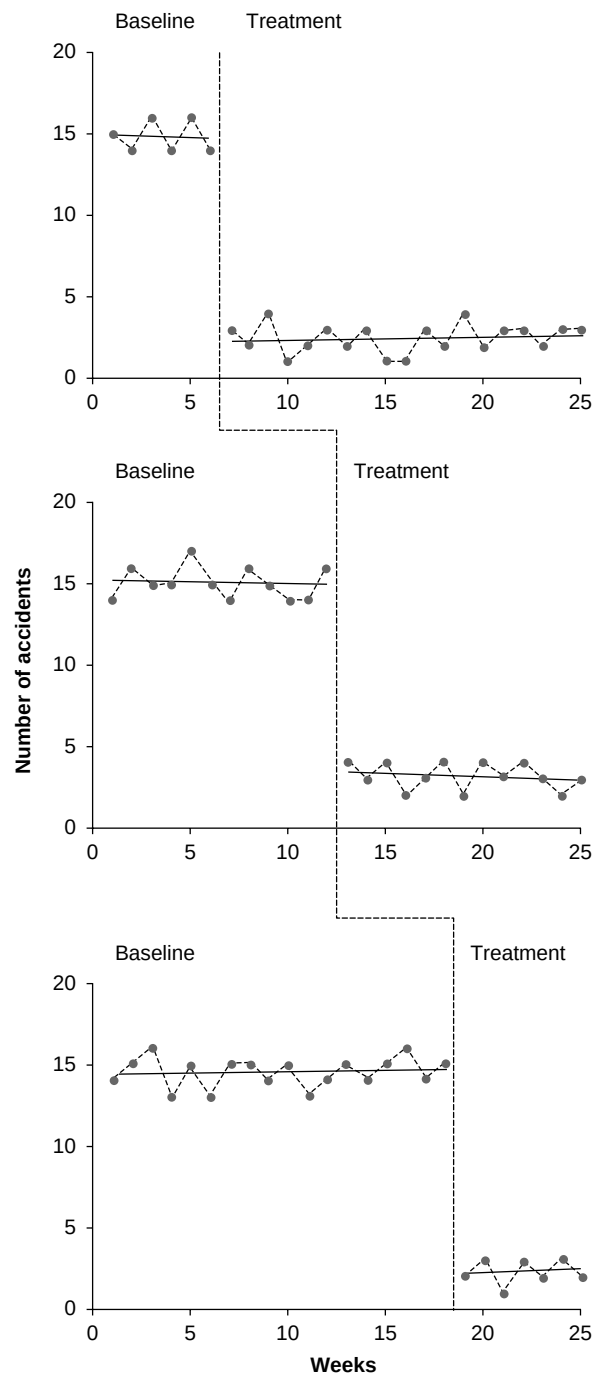


Figure 2 A Multiple-Baseline Design Across Settings

Note: Baseline data are collected on all baselines, and the intervention is introduced in a staggered manner. As with the withdrawal design, the baseline data provide a predicted level of the dependent variable in absence of the treatment. This figure shows an effective intervention in that the level of the dependent variable (the number of accidents) decreases when the treatment is applied. We can rule out coincidental variables as causes for the change because the dependent variable changes only when the intervention is applied and not at other times.

depicts a multiple-baseline design across settings using the same details as in Figure 1. Each baseline represents a different job site.

Unlike withdrawal designs, multiple-baseline designs can be used to study irreversible treatments and treatments that, because of their effectiveness, would be undesirable to withdraw. Nevertheless, multiple-baseline designs have the following shortcomings: (a) They require the withholding of treatment for extended periods of time, which may be undesirable in some cases; (b) they require that the DVs measured across the different baselines be independent of each other; and (c) they may yield ambiguous data regarding treatment effects when changes occur during some treatment phases and not during others.

—Bradley E. Huitema
and Sean Laraway

Further Reading

- Barlow, D. H., & Hersen, M. (1984). *Single case experimental design: Strategies for studying behavior change* (2nd ed.). Needham Heights, MA: Allyn & Bacon.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- Kazdin, A. E. (1998). The case study and single-case research designs. In A. E. Kazdin (Ed.), *Research design in clinical psychology* (3rd ed., pp. 202–244). Needham Heights, MA: Allyn & Bacon.
- Poling, A., Methot, L. L., & LeSage, M. G. (1995). *Fundamentals of behavior analytic research*. New York: Plenum Press.
- Sidman, M. (1960). *Tactics of scientific research*. Boston: Publisher's Cooperative.

Journal of Applied Behavior Analysis: <http://seab.envmed.rochester.edu/jaba/index.html>

SINGULAR AND GENERALIZED SINGULAR VALUE DECOMPOSITION

The singular value decomposition (SVD), a generalization of the eigendecomposition, can be used to

analyze rectangular matrices (the eigendecomposition is defined only for squared matrices). By analogy with the eigendecomposition, which decomposes a matrix into two simple matrices, the main idea of the SVD is to decompose a rectangular matrix into three simple matrices: two orthogonal matrices and one diagonal matrix.

Because it gives a least square estimate of a given matrix by a lower rank matrix of the same dimensions, the SVD is equivalent to principal components analysis (PCA) and metric multidimensional scaling and is therefore an essential tool for multivariate analysis. The generalized SVD (GSVD) decomposes a rectangular matrix and takes into account constraints imposed on the rows and the columns of the matrix. The GSVD gives a weighted generalized least square estimate of a given matrix by a lower rank matrix, and therefore, with an adequate choice of the constraints, the GSVD implements all linear multivariate techniques (e.g., canonical correlation, linear discriminant analysis, and correspondence analysis).

Definitions and Notations

Recall that a *positive semidefinite* matrix can be obtained as the product of a matrix by its transpose. This matrix is obviously square and symmetric, and (less obviously) its eigenvalues are all positive or null, and the eigenvectors corresponding to different eigenvalues are pairwise orthogonal. Let $\mathbf{X}$ be a positive semidefinite matrix; then its eigendecomposition is expressed as

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad (1)$$

with $\mathbf{U}$ being an orthonormal matrix (i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$) and $\mathbf{\Lambda}$ being a diagonal matrix containing the eigenvalues of $\mathbf{X}$.

The SVD uses the eigendecomposition of a positive semidefinite matrix in order to derive a similar decomposition applicable to all rectangular matrices composed of real numbers. The idea here is to decompose any matrix into three simple matrices, two orthonormal matrices and one diagonal matrix. When applied

to a positive semidefinite matrix, the SVD is equivalent to the eigendecomposition.

Formally, if $\mathbf{A}$ is a rectangular matrix, its SVD decomposes it as

$$\mathbf{A} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T, \quad (2)$$

with

$\mathbf{P}$ is the (normalized) eigenvectors of the matrix $\mathbf{A}\mathbf{A}^T$ (i.e., $\mathbf{P}^T\mathbf{P} = \mathbf{I}$). The columns of $\mathbf{P}$ are called the *left singular vectors* of $\mathbf{A}$.

$\mathbf{Q}$ is the (normalized) eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$ (i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$). The columns of $\mathbf{Q}$ are called the *right singular vectors* of $\mathbf{A}$.

$\mathbf{\Delta}$ is the diagonal matrix of the *singular values*, $\mathbf{\Delta} = \mathbf{\Lambda}^{\frac{1}{2}}$, with $\mathbf{\Lambda}$ being the diagonal matrix of the eigenvalues of matrix $\mathbf{A}\mathbf{A}^T$ and of the matrix $\mathbf{A}^T\mathbf{A}$ (they are the same).

The SVD is a consequence of the eigendecomposition of a positive semidefinite matrix. This can be shown by considering the eigendecomposition of the two positive semidefinite matrices that can be obtained from $\mathbf{A}$: namely $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. If we express these matrices in terms of the SVD of $\mathbf{A}$, we obtain the following equations:

$$\mathbf{A}\mathbf{A}^T = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T\mathbf{Q}\mathbf{\Delta}\mathbf{P}^T = \mathbf{P}\mathbf{\Delta}^2\mathbf{P}^T = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T, \quad (3)$$

and

$$\mathbf{A}^T\mathbf{A} = \mathbf{Q}\mathbf{\Delta}\mathbf{P}^T\mathbf{P}\mathbf{\Delta}\mathbf{Q}^T = \mathbf{Q}\mathbf{\Delta}^2\mathbf{Q}^T = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T. \quad (4)$$

This shows that $\mathbf{\Delta}$ is the square root of $\mathbf{\Lambda}$, that $\mathbf{P}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$, and that $\mathbf{Q}$ are the eigenvectors of $\mathbf{A}^T\mathbf{A}$.

For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix} \quad (5)$$

can be expressed as

$$\begin{aligned} \mathbf{A} &= \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \\ &= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & 0.7071 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix}. \end{aligned} \quad (6)$$

We can check that

$$\begin{aligned} \mathbf{A}\mathbf{A}^T &= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & 0.7071 \end{bmatrix} \begin{bmatrix} 2^2 & 0 \\ 0 & 1^2 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.8165 & -0.4082 & -0.4082 \\ 0 & -0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 2.6667 & -1.3333 & -1.3333 \\ -1.3333 & 1.1667 & 0.1667 \\ -1.3333 & 0.1667 & 1.1667 \end{bmatrix} \end{aligned} \quad (7)$$

and that

$$\begin{aligned} \mathbf{A}^T\mathbf{A} &= \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} 2^2 & 0 \\ 0 & 1^2 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 2.5 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}. \end{aligned} \quad (8)$$

Technical Note: Agreement Between Signs

Singular vectors come in pairs made of one left and one right singular vector corresponding to the same singular value. They could be computed

separately or as a pair. Equation 2 requires computing the eigendecomposition of two matrices. Rewriting this equation shows that it is possible, in fact, to compute only one eigendecomposition. As an additional bonus, computing only one eigendecomposition prevents a problem that can arise when the singular vectors are obtained from two separate eigendecompositions. This problem follows from the fact that the eigenvectors of a matrix are determined up to a multiplication by -1 but that singular vectors, being pairs of eigenvectors, need to have compatible parities. Therefore, when computed as eigenvectors, a pair of singular vectors can fail to reconstruct the original matrix because of this parity problem.

This problem is illustrated by the following example: The matrix

$$\mathbf{A} = \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix} \quad (9)$$

can be decomposed in two equivalent ways:

$$\begin{aligned} \mathbf{A} &= \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \\ &= \begin{bmatrix} 0.8165 & 0 \\ -0.4083 & -0.7071 \\ -0.4083 & 0.7071 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} -0.8165 & 0 \\ 0.4083 & 0.7071 \\ 0.4083 & 0.7071 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad (10) \\ &\quad \begin{bmatrix} -0.7071 & -0.7071 \\ 0.7071 & -0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix}. \end{aligned}$$

But when the parity of the singular vectors does not match, the SVD will fail to reconstruct the original matrix, as illustrated by

$$\begin{aligned} \mathbf{A} &\neq \begin{bmatrix} -0.8165 & 0 \\ 0.4083 & 0.7071 \\ 0.4083 & 0.7071 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \quad (11) \\ &= \begin{bmatrix} -1.1547 & -1.1547 \\ 1.0774 & 0.0774 \\ 0.0774 & 1.0774 \end{bmatrix}. \end{aligned}$$

By computing only one matrix of singular vectors, we can rewrite Equation 2 in a manner that expresses that one matrix of singular vectors can be obtained from the other:

$$\mathbf{A} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \iff \mathbf{P} = \mathbf{A}\mathbf{Q}\mathbf{\Delta}^{-1}. \quad (12)$$

For example:

$$\begin{aligned} \mathbf{P} &= \mathbf{A}\mathbf{Q}\mathbf{\Delta}^{-1} \\ &= \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \quad (13) \\ &= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & 0.7071 \end{bmatrix}. \end{aligned}$$

Generalized Singular Value Decomposition

For a given $I \times J$ matrix $\mathbf{A}$, generalizing SVD involves using two positive definite square matrices with sizes $I \times I$ and $J \times J$, respectively. These two

matrices express constraints imposed on the rows and the columns of $\mathbf{A}$, respectively. Formally, if $\mathbf{M}$ is the $I \times I$ matrix expressing the constraints for the rows of $\mathbf{A}$ and $\mathbf{W}$ the $J \times J$ matrix of the constraints for the columns of $\mathbf{A}$, then the matrix $\mathbf{A}$ is now decomposed into

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{V}}^T \text{ with: } \tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}} = \tilde{\mathbf{V}}^T\mathbf{W}\tilde{\mathbf{V}} = \mathbf{I}. \quad (14)$$

In other words, the generalized singular vectors are orthogonal under the constraints imposed by $\mathbf{M}$ and $\mathbf{W}$.

This decomposition is obtained as a result of the standard singular value decomposition. We begin by defining the matrix $\tilde{\mathbf{A}}$ as:

$$\tilde{\mathbf{A}} = \mathbf{M}^{\frac{1}{2}}\mathbf{A}\mathbf{W}^{\frac{1}{2}} \Leftrightarrow \mathbf{A} = \mathbf{M}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{W}^{-\frac{1}{2}}. \quad (15)$$

We then compute the standard singular value decomposition as $\tilde{\mathbf{A}}$:

$$\tilde{\mathbf{A}} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \text{ with: } \mathbf{P}^T\mathbf{P} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}. \quad (16)$$

The matrices of the generalized eigenvectors are obtained as

$$\tilde{\mathbf{U}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{P} \text{ and } \tilde{\mathbf{V}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{Q}. \quad (17)$$

The diagonal matrix of singular values is simply equal to the matrix of singular values of $\tilde{\mathbf{A}}$:

$$\tilde{\mathbf{\Delta}} = \mathbf{\Delta} \quad (18)$$

We verify that

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{V}}^T$$

by substitution:

$$\begin{aligned} \mathbf{A} &= \mathbf{M}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{W}^{-\frac{1}{2}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{P}\mathbf{\Delta}\mathbf{Q}^T\mathbf{W}^{-\frac{1}{2}} \\ &= \tilde{\mathbf{U}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{V}}^T \text{ (from Equation 17).} \end{aligned} \quad (19)$$

To show that Equation 14 holds, suffice it to show that

$$\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}} = \mathbf{P}^T\mathbf{M}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}^{-\frac{1}{2}}\mathbf{P} = \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad (20)$$

and

$$\tilde{\mathbf{V}}^T\mathbf{W}\tilde{\mathbf{V}} = \mathbf{Q}^T\mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \quad (21)$$

Mathematical Properties

It can be shown that the SVD has the important property of giving an optimal approximation of a matrix by another matrix of smaller rank. In particular, the SVD gives the best approximation, in a least square sense, of any rectangular matrix by another rectangular of the same dimensions but smaller rank.

Precisely, if $\mathbf{A}$ is an $I \times J$ matrix of rank L (i.e., $\mathbf{A}$ contains L singular values that are not zero), we denote by $\mathbf{P}_{[K]}$ (respectively $\mathbf{Q}_{[K]}$, $\mathbf{\Delta}_{[K]}$) the matrix made of the first K columns of $\mathbf{P}$ (respectively $\mathbf{Q}$, $\mathbf{\Delta}$):

$$\mathbf{P}_{[K]} = [\mathbf{p}_1, \dots, \mathbf{p}_K] \quad (22)$$

$$\mathbf{Q}_{[K]} = [\mathbf{q}_1, \dots, \mathbf{q}_K] \quad (23)$$

$$\mathbf{\Delta}_{[K]} = \text{diag}\{\delta_1, \dots, \delta_K\} \quad (24)$$

The matrix $\mathbf{A}$ reconstructed from the first K eigenvectors is denoted $\mathbf{A}_{[K]}$. It is obtained as

$$\mathbf{A}_{[K]} = \mathbf{P}_{[K]}\mathbf{\Delta}_{[K]}\mathbf{Q}_{[K]}^T = \sum_k^K \delta_k \mathbf{p}_k \mathbf{q}_k^T \quad (25)$$

(with δ_k being the k th singular value).

The reconstructed matrix $\mathbf{A}_{[K]}$ is said to be optimal (in a least square sense) for matrices of rank K because it satisfies the following condition:

$$\begin{aligned} \|\mathbf{A} - \mathbf{A}_{[K]}\|^2 &= \text{trace} \left\{ (\mathbf{A} - \mathbf{A}_{[K]}) (\mathbf{A} - \mathbf{A}_{[K]})^T \right\} \\ &= \min_{\mathbf{X}} \|\mathbf{A} - \mathbf{X}\|^2 \end{aligned} \quad (26)$$

for the set of matrices $\mathbf{X}$ of rank smaller than or equal to K . The quality of the reconstruction is given by the ratio of the first K eigenvalues (i.e., the squared singular values) to the sum of all the eigenvalues. This quantity is interpreted as *the reconstructed proportion* or *the explained variance*; it corresponds to the inverse of the quantity minimized by Equation 26. The quality of reconstruction can also be interpreted as the squared coefficient of correlation (precisely as the R_v coefficient, see entry) between the original matrix and its approximation.

The GSVD minimizes an expression similar to Equation 26, namely,

$$\mathbf{A}_{[K]} = \min_{\mathbf{X}} [\text{trace} \{ \mathbf{M}(\mathbf{A} - \mathbf{X})\mathbf{W}(\mathbf{A} - \mathbf{X})^T \}], \quad (27)$$

for the set of matrices $\mathbf{X}$ of rank smaller than or equal to K .

SVD and General Linear Model

It can be shown that the SVD of a rectangular matrix gives this PCA of the matrix, with, for example, the factor scores being obtained as $\mathbf{F} = \mathbf{P}\mathbf{\Delta}$.

The adequate choice of matrices $\mathbf{M}$ and $\mathbf{W}$ makes the GSVD a very versatile tool that can implement the set of methods of linear multivariate analysis. For example, correspondence analysis (see entry) can be implemented by using a probability matrix (i.e., made of positive or null numbers whose sum is equal to 1) along with two diagonal matrices, $\mathbf{M} = \mathbf{D}_r$ and $\mathbf{W} = \mathbf{D}_c$, representing the relative frequencies of the rows and the columns of the data matrix, respectively. The other multivariate techniques (e.g., discriminant analysis and canonical correlation analysis) can be implemented with the proper choice of the matrices $\mathbf{M}$ and $\mathbf{W}$.

An Example of Singular Value Decomposition: Image Compression

The SVD of a matrix is equivalent to PCA. We illustrate this property by showing how it can be used to perform image compression. Modern technology uses digitized pictures, which are equivalent to a matrix giving the gray level value of each pixel. For example, the matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 0 & 2 & 2 & 2 & 0 & 2 & 2 & 2 & 0 & 2 & 2 & 2 & 1 \\ 1 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 1 \\ 1 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 1 \\ 1 & 0 & 2 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad (28)$$

corresponds to the image in Figure 1.

In general, pictures coming from natural images have rather large dimensions. For example, the picture shown in Figure 2 corresponds to a matrix with 204 rows and 290 columns (therefore $204 \times 290 = 59,160$ pixels). To avoid the problem of



Figure 1 The Matrix of Equation 28 Displayed as a Picture

transmitting or storing the numerical values of such large images, we want to represent the image with fewer numerical values than the original number of pixels.

Thus, one way of compressing an image is to compute the singular value decomposition and then to reconstruct the image by an approximation of smaller rank. This technique is illustrated in Figures 4 and 5, which show the terms $\mathbf{p}_k \mathbf{q}_k^T$ and the terms $\sum \mathbf{p}_k \mathbf{q}_k^T$, respectively. As can be seen in Figure 4, the image is reconstructed almost perfectly (according to the human eye) by a rank 25 approximation. This gives a compression ratio of

$$\begin{aligned} 1 - \frac{25 \times (1 + 204 + 290)}{204 \times 290} \\ = 1 - .2092 = .7908 \approx 80\%. \end{aligned} \quad (29)$$

—Hervé Abdi



Figure 2 A Picture Corresponding to a Matrix on the Order of $204 \times 290 = 59,160$ Pixels



Figure 3 The Picture in Figure 2 Built Back With 25 Pairs of Singular Vectors (Compression Rate of $\approx 80\%$)



Figure 4 Reconstruction of the Image in Figure 2

Note: The percentages of explained variance are 0.9347; 0.9512; 0.9641; 0.9748; 0.9792; 0.9824; 0.9846; 0.9866; 0.9881; 0.9896; 0.9905; 0.9913; 0.9920; 0.9926; 0.9931; 0.9936; 0.9940; 0.9944; 0.9947; 0.9950; 0.9953; 0.9956; 0.9958; 0.9961; 0.9963.

See also Correspondence Analysis; Discriminant Analysis; Discriminant Correspondence Analysis; DISTATIS; Eigendecomposition; Least Squares, Method of; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Multiple Factor Analysis; Principal Component Analysis; R_v and Congruence Coefficients; STATIS

Further Reading

Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.

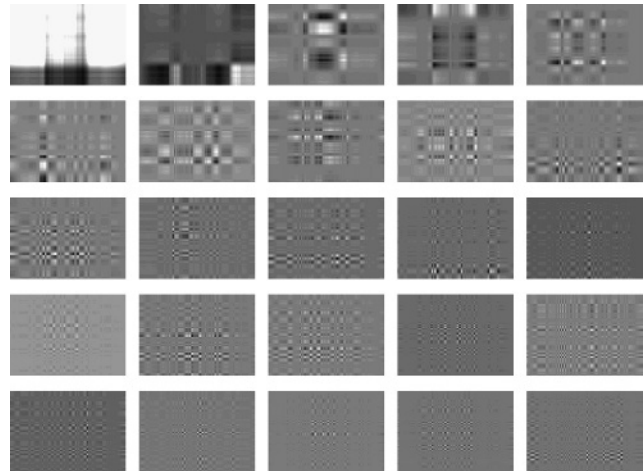


Figure 5 The Terms $p_k q_k$ Used to Reconstruct the Image in Figure 2 (see Figure 4)

Note: The eigenvalues (squared singular values) associated with each image are 0.9347; 0.0166; 0.0129; 0.0107; 0.0044; 0.0032; 0.0022; 0.0020; 0.0015; 0.0014; 0.0010; 0.0008; 0.0007; 0.0006; 0.0005; 0.0005; 0.0004; 0.0004; 0.0003; 0.0003; 0.0003; 0.0003; 0.0003; 0.0003; 0.0002; 0.0002; 0.0002.

Abdi, H., & Valentin, D. (2006). *Mathématiques pour les sciences cognitives* (Mathematics for cognitive sciences). Grenoble, France: Presses Universitaires de Grenoble.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Strang, G. (2003). *Introduction to linear algebra*. Cambridge, MA: Wellesley-Cambridge Press.

SIX SIGMA

The term *Six Sigma* may be interpreted in several ways. To some, it represents a metric denoting the level of quality of a product or process, such as the proportion of nonconforming product. This may be expressed as a proportion or percentage (5% nonconforming product, say) or in parts per million (ppm), such as 5,000 defects per million opportunities. To others, such as senior management, Six Sigma conveys a philosophy of continuous quality improvement. This is akin to accepting Six Sigma as the organization motto, with everyone committed to improving products, processes, and services along such dimensions as quality, price, lead time, and on-time delivery. It is interpreted as a business strategy. Finally, Six Sigma is viewed as a methodology for improving quality. In

this context, it represents a structured approach to problem solving using a variety of statistical and interventional tools and techniques.

Perspective on Six Sigma as a Metric

Six Sigma was coined by Motorola in the 1980s. The Greek letter sigma (σ) represents the variability in a process as measured by its standard deviation. The conceptual perspective of the term Six Sigma is that it is desirable for a process to be operating such that the *specification limits*, which govern customer requirements on a selected quality characteristic, are located 6 standard deviations away from the mean value of the characteristic. It is assumed that the probability distribution of the quality characteristic (X) is normal, and its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x - \mu)^2/2\sigma^2),$$

where μ and σ represent the process mean and standard deviation, respectively. When the assumption of normality of distribution of the quality characteristic can be justified, using standard normal tables, it can be shown that the proportion of nonconforming product is 0.002 ppm. The tail area under the normal curve, above or below the appropriate specification limit, is 0.001 ppm on each side. Figure 1 demonstrates this concept.

The above interpretation of Six Sigma as a metric applies in a static situation. However, Motorola wanted to account for drifts in the process in the long

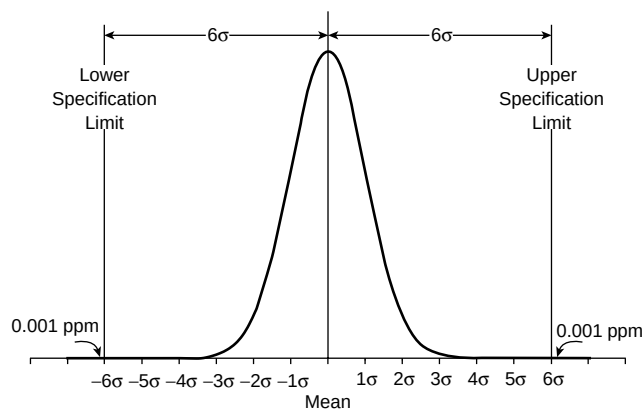


Figure 1 Process Distribution With Specification Limits at 6 Standard Deviations From Mean

run. An assumption was made that the drift in the process mean would be no more than 1.5 standard deviations from its original location. This assumption implies that a system of statistical process control will be in place for detecting when a process goes “out of control.”

A process could go out of control because of special causes. These special causes are not part of the system as designed but could occur as a result of the use of a wrong tool, an improper raw material, or an operator error. Identification of special causes is typically accomplished through the use of control charts. Special causes could cause a shift in the process mean or an increase in the process standard deviation. Common causes, on the other hand, are inherent to the system. A process can be redesigned to reduce the impact of the common causes, but they can never be eliminated completely. A process with only common causes is said to be in statistical control.

Motorola made the assumption that if a process mean drifted more than 1.5 standard deviations from its initial location, statistical process control methods should be able to detect the drift and bring the process back to a state of control. Thus, at the worst, if the process mean drifts 1.5 standard deviations and the drift is not detected, the process mean would be 4.5 standard deviations from the closer specification limit and 7.5 standard deviations from the farther specification limit. The premise is that the process mean was initially centered between the specification limits, with each limit being 6 standard deviations away from the mean. Under the assumption of normality, it can be shown that the proportion of nonconforming product outside the closer specification limit is about 3.4 ppm, and the proportion nonconforming outside the farther specification limit is negligible. Thus, the total proportion of nonconforming product is 3.4 ppm. Figure 2 demonstrates this concept; it is assumed that the shift in the process mean is 1.5 standard deviations in either direction.

Almost all the Six Sigma literature, when viewed in the context of a metric, highlights this nonconformance rate of 3.4 ppm. As explained in Figure 2, the process mean, in this situation, is really 4.5 standard deviations from the closer specification limit and not 6 standard deviations, as in the original condition. Perhaps an appropriate term to denote this situation

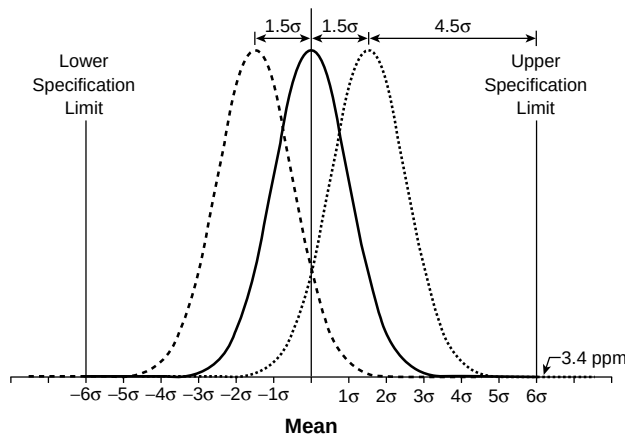


Figure 2 Process Distribution With a Shift of 1.5 Standard Deviations From Mean

would be Four and a Half Sigma, which would be consistent with a nonconformance rate of 3.4 ppm.

Inherent Assumptions and Their Validity

The use of Six Sigma as a metric includes two inherent assumptions. First, it is assumed that the distribution of the quality characteristic, which is being compared to a set of defined specification limits, is normal. There is no guarantee that this assumption will be satisfied for all quality characteristics. The proportion nonconforming will be influenced not only by the location of the process mean and the magnitude of the process variability but also by the shape of the distribution. For example, suppose a product quality characteristic is found to follow an exponential distribution with parameter λ . The probability density function is

$$f(x) = \lambda e^{-\lambda x},$$

where the mean and standard deviation are both $1/\lambda$. Suppose a lower specification limit (LSL) is specified. The proportion nonconforming is

$$P[X < \text{LSL}] = 1 - \exp((-1/\lambda) \text{LSL}).$$

In practice, λ may be estimated from historical data and replaced by its estimate, $\hat{\lambda}$. If so, the proportion nonconforming found from the above equation will

not necessarily be equal to that obtained using the normal probability distribution. Thus, the metric of 3.4 ppm nonconformance for a Six Sigma process may not necessarily be correct, even if the closest specification limit is 4.5 standard deviations from the process mean.

The second assumption deals with the process shift of 1.5 standard deviations from the original location, assuming that the specifications were initially 6 standard deviations from the process mean. Motorola could have made this assumption. Certainly, for all types of products or services in various industries, there is no guarantee that such a degree of process shift takes place. Nor is there evidence that, as a rule, shifts of 1.5 standard deviations will be detected and the process mean restored through remedial action to its original location. It is assumed that shifts greater than 1.5 standard deviations will not happen, but this level cannot be guaranteed.

Further, there might be instances whereby, based on the criticality of the components or the product quality characteristic, it is not desirable to allow a shift of 1.5 standard deviations; detection might need to take place sooner. In addition, on mathematical grounds, there is no justification that for all processes, shifts of up to 1.5 standard deviations may take place prior to detection.

Six Sigma as a Methodology

One of the major contributions of Six Sigma lies in its application as a structured approach to problem solving and improvements. When applied to improving processes (manufacturing, transactional, or service), Six Sigma provides a road map. There are five main phases in this approach: Define, Measure, Analyze, Improve, and Control.

The Define phase includes problem statement and identification of the scope of the project. It may also include looking at high-level process maps and defining the metrics to be used, as well as selection of baseline and improvement goals. Implementation of Six Sigma is typically accomplished on a project-by-project basis. The person who serves as the promoter of the particular project is called a *champion* and is

also a liaison to senior management. Top management support is enhanced through adequate interaction with the champion, who conveys progress and associated project needs to senior administration.

The Measure phase involves identifying customer needs and documenting the process. Customer requirements are prioritized and measures are established with a tool known as *quality function deployment*. Key process output variables that impact process performance are noted, and associated key process input variables that can be monitored are identified. All steps involved in the process are placed on a process map. Here, for each activity, inputs (controllable and noise), outputs, ownership, and possible reasons for failure are clearly delineated. Metrics for determining process performance levels are also developed. They may include such measures as *defects per million opportunities* or sigma level, which is an indication of the location of the nearest specification limit relative to the process mean. The larger the sigma level, the better. Other measures include short-term process capability indices such as C_p , C_{pk} , CPL , and CPU . C_p is a measure of process potential and should be used when the process mean, μ , is centered between specification limits. It is given by the relation

$$C_p = \frac{USL - LSL}{6\sigma},$$

where USL and LSL denote the upper specification limit and lower specification limit, respectively, and σ represents the process standard deviation. Note that, assuming normality of distribution, the denominator, 6σ , represents the process spread. This is the distance between the *upper natural tolerance limit* and *lower natural tolerance limit*, each of which is 3 standard deviations from the process mean. Hence, C_p can be interpreted as the specification spread divided by the process spread. Desirable values of C_p are greater than or equal to 1. Figure 3 demonstrates a capable process in which $C_p > 1$.

A drawback of C_p is that it does not consider the location of the process. Thus, even if the process mean were to be close to the upper specification limit and the process variability remained the same as that in Figure 3, the C_p index would still be greater than 1, indicating a

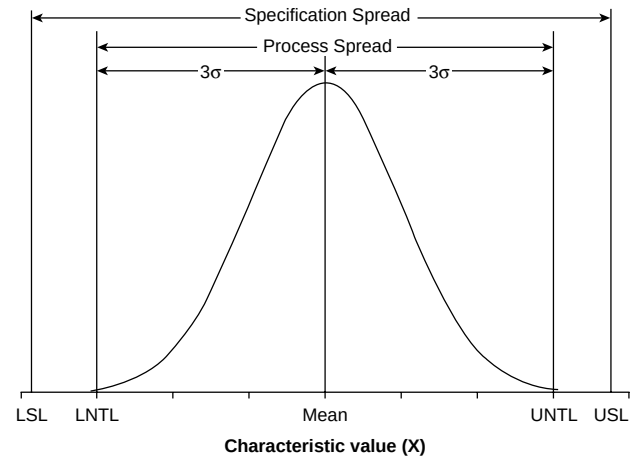


Figure 3 A Capable Process With $C_p > 1$

desirable process. However, in this situation, there would be quite a bit of nonconforming product.

In some cases, specifications are one-sided only. For example, the tensile strength of a cable must meet at least a lower specification limit, or the waiting time before being served by a teller in a bank must be less than an upper specification limit. In these instances, the short-term capability indices for one-sided specifications are given by

$$CPL = \frac{\mu - LSL}{3\sigma} \text{ or}$$

$$CPU = \frac{USL - \mu}{3\sigma},$$

respectively. As before, desirable values of CPL or CPU are greater than or equal to 1.

To overcome the drawback of the C_p index, the C_{pk} index, which uses information on both process mean and standard deviation, is used. It is given by

$$C_{pk} = \min \left[\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma} \right] \\ = \min [CPU, CPL].$$

It is observed that C_{pk} represents the scaled distance, relative to 3 standard deviations, between the process mean and the closest specification limit. Desirable values of C_{pk} are $C_{pk} \geq 1$. Figure 4 shows a

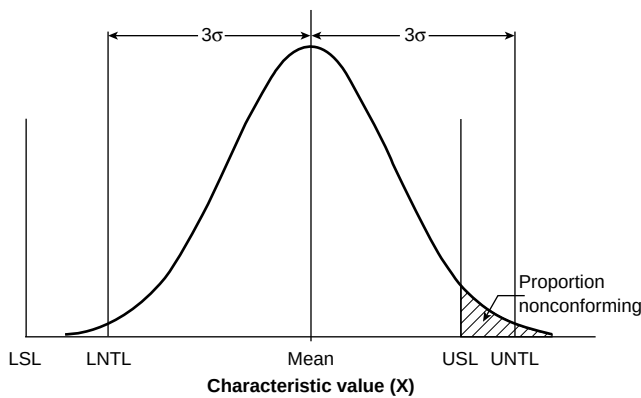


Figure 4 A Process That Is Not Capable, With $C_{pk} < 1$

process that is not capable, with nonconforming product above the upper specification limit. The C_{pk} value for this process is < 1 .

In the long term, because a process can drift, an estimate of the process standard is calculated by using data from subgroups of information collected over a time span. This estimate is denoted by σ_{LT} , as opposed to the estimate of the short-term process standard (σ_{ST}), which is estimated from the variation within subgroups. Thus, there are equivalent long-term versions for all the previous process capability indices. These are denoted P_p , PPL , PPU , and P_{pk} , with the only difference in the computation being that the short-term estimate of the process standard deviation is replaced by a long-term estimate.

Also in the Measure phase, capability analysis of the measurement system is conducted to determine its stability. Such studies are known as *gauge* repeatability and reproducibility studies. Repeatability refers to the inherent variation present in the gage system, implying the variation in measured values when the same part is measured by the same inspector using the same gage, over and over again. Reproducibility, on the other hand, refers to the variability between inspectors or the interaction between inspectors and parts.

The Analyze phase consists of identifying root causes of variation by collecting data from the process and analyzing it with tools such as cause-and-effect or fishbone diagrams. The tools may also include histograms, box plots, scatter diagrams, run charts, multivariable plots, and distribution plots. Distribution

plots, such as the normal probability plot, are used to check the validity of the distributional assumptions. Descriptive statistics on the process-related variables, such as mean, standard deviation, skewness coefficient, and kurtosis coefficient, are commonly used. Inferential statistics is also a focal point of analysis and may include constructing confidence intervals and testing hypotheses on population parameters such as population means, proportions, variances, difference of two means, difference of two proportions, or ratio of two variances. In situations involving comparison of more than two population means, analysis of variance is used. Also, when it is desired to develop a statistical model for a chosen response variable based on certain independent variables or predictors, regression analysis may be used. Software packages can perform all these analyses.

The Improve phase focuses on how process performance may be elevated to reach certain goals not attainable by the current process. Design-of-experiments concepts are used. This phase may start with process characterization, in which key factors or process variables that impact the response variable are identified. Screening designs, such as two-level factorial or two-level fractional experiments, or Plackett-Burman designs may be used in this context. Once the vital factors are identified, the next step is process optimization. Here it is of interest to locate the region in the important factors that leads to the best possible response. The objective is to determine the optimal settings of the factors such that a target value of the response is achieved. Factorial designs or response surface designs are possibilities, with central composite designs or Box-Behnken designs being alternatives to explore when considering polynomial models. Usually, analysis of variance procedures are used for the determination of statistical significance of the factors. The last stage in this phase is the verification or validation step. Experiments are conducted at the identified optimal factor settings, and the observed response variable values are analyzed to determine their proximity to the predicted values.

The final phase of the Six Sigma process is the Control phase. The aim is to sustain the gains identified in the previous phase. Opportunities for standardizing

process steps and mistake proofing the process could be identified. Statistical process control is accomplished with control charts, which detect the presence of special causes in the system and take remedial action as soon as possible. Rules exist for triggering the detection of special causes, the most common one being a single point falling outside the control limits. Control charts exist for variables as well as for attributes; charts for the mean and range, mean and standard deviation, individuals, and moving range are common. Refinements include the exponentially weighted moving average, cumulative sum, and control charts for short production runs. Under attribute charts, charts for proportion nonconforming, number nonconforming, number of defects or nonconformities, and number of nonconformities per unit are commonly used.

Conclusions

While the majority of the statistical or nonstatistical tools used in the Six Sigma approach are not new, there are a few unique features. First, Six Sigma is associated with an approach oriented toward bottom-line results. Financial impact of the project is of major concern, as is retention of top management commitment. Second, Six Sigma identifies a business strategy that may be used for continual improvement to keep abreast of competitors. Third, it utilizes a structured approach to improve a process, product, or services. Such an approach facilitates a better understanding of critical business processes. Fourth, it maintains and improves communication between operations and management through project presentations. Fifth, it demonstrates the value of the various tools in the organizational context, thus creating a win-win situation for everyone.

—Amitava Mitra

Further Reading

Breyfogle, F. W., III, Cupello, J. M., & Meadows, B. (2001). *Managing Six Sigma: A practical guide to understanding, assessing, and implementing the strategy that yields bottom-line success*. New York: Wiley.

Pyzdek, T. (2001). *The Six Sigma handbook*. New York: McGraw-Hill.

Six Sigma article: http://en.wikipedia.org/wiki/Six_Sigma

SIXTEEN PERSONALITY FACTOR QUESTIONNAIRE

The Sixteen Personality Factor (16PF) Questionnaire (published by the Institute of Personality and Ability Testing, www.ipat.com) is a comprehensive self-report inventory of normal personality that takes about 35 minutes to complete. Unlike most test authors, who try to assess a particular psychological construct, Raymond B. Cattell designed the 16PF to measure personality itself. In the 1940s, Cattell defined personality as anything about people that allows us to predict how they will behave in a given situation. He believed that any normal personality trait would be reflected in the English language, as people need adjectives to describe each other. He studied how adjectives clung together statistically as people used them to describe themselves and others. A factor analysis of the results produced the 16 factors; these, in theory, account for most of what is meant by personality.

The 16 factors are Warmth, Reasoning, Emotional Stability, Dominance, Liveliness, Rule-Consciousness, Social Boldness, Sensitivity, Vigilance, Abstractedness, Privateness, Apprehension, Openness to Change, Self-Reliance, Perfectionism, and Tension. Cattell wrote test items to measure these factors. For example, the item “I would rather be a seal in a seal colony than an eagle on a cliff” measured the first factor, Warmth, or sociability. Cattell considered intelligence to be a personality factor because it helps predict behavior, and the 16PF contains some word problems to measure reasoning. The test also has scales that measure the participant’s response style. The latest version of the test, the fifth edition, was published in 1993.

After developing the test and collecting profiles from numerous people, Cattell factor-analyzed 16PF

results. Not counting intelligence, five factors emerged, which the Institute of Personality and Ability Testing has named Extraversion, Anxiety, Self-Control, Independence, and Tough-Mindedness. Thus, Cattell started the Big Five theory, but he always thought that five factors were too few to describe human personality.

As a test of normal personality, the 16PF-5 is well suited to industrial uses, and it can be used under the Americans with Disabilities Act for preemployment screening. With its excellent psychometrics, it has become one of the most widely used personality tests in the world, available in nearly 40 languages. As psychotherapy became destigmatized in the 20th century, and as normal people seek help more than ever before, the 16PF has been increasingly used in the clinic as well.

—Michael Karson

See also Personality Tests

Further Reading

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.

Karson, M., Karson, S., & O'Dell, J. (1997). *16PF interpretation in clinical practice: A guide to the fifth edition*. Champaign, IL: Institute for Personality and Ability Testing.

Institute for Personality and Ability Testing information about the test, its underlying theory, and interpretive programs: www.ipat.com

SKEWNESS

Skewness is a measure of the lack of symmetry, or the lopsidedness, a distribution has. In other words, one tail of the distribution is longer than another. A positively skewed distribution has a longer right tail than left, corresponding to a smaller number of occurrences at the high end of the distribution. This might be the case when you have a test that is very difficult: Few people get scores that are very high, and many more get scores that are relatively low. A negatively

skewed distribution has a shorter right tail than left, corresponding to a larger number of occurrences at the high end of the distribution. This would be the case for an easy test (lots of high scores and relatively few low scores).

Although skewness (and kurtosis) are used mostly as descriptive terms (such as “That distribution is negatively skewed”), mathematical indicators can indicate how skewed or kurtotic a distribution is. For example, skewness is computed by subtracting the value of the median from the mean:

$$Sk = \frac{3(\bar{X} - M)}{s},$$

where

Sk is Pearson measure of skewness,

$\bar{X}$ is the mean,

M is the median, and

s is the standard deviation.

For example, if the mean of a distribution is 100 and the median is 95, the skewness value is $100 - 95 = 5$, and the distribution is positively skewed. If the mean of a distribution is 85 and the median is 90, the skewness value is $85 - 90 = -5$, and the distribution is negatively skewed. The formula takes the standard deviation of the distribution into account so that skewness indicators can be compared with one another.

—Neil J. Salkind

See also Kurtosis

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

SMOOTHING

Smoothing methods attempt to capture the underlying structure of data that contain noise. Noise in data may result from measurement imprecision or the effect of

unmeasured variables, and noise tends to mask the structure of a data set or the relationship between two variables. Smoothers aim to eliminate this noise while making very few model assumptions about the distribution of the data.

Common smoothing techniques include density estimation and (nonparametric) regression curve estimation. Density estimation uses sample data for a single variable to estimate the population distribution of that variable. While parametric models (e.g., normal distribution, exponential distribution) can be fit for such data, smoothers typically assume only that the variable is continuous with a smooth density function. Nonparametric regression models the relationship between two (or more) variables without assuming a specific functional form (such as linear or quadratic) for the regression curve.

The smooth curve (whether a density estimate or regression curve) is often determined by “local weighting”—that is, the curve value at any point on the graph is a weighted average of the observed data values near that point, with data closest to the point receiving the greatest weight. Most smoothing methods incorporate some type of tuning parameter that allows the user to control the smoothness of the estimated curve.

This following example (Figure 1) relates the top speed and gas mileage of 82 car models using data from the U.S. Environmental Protection Agency and available online at <http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html>.

The smooth curve shows the negative association between speed and mileage, reflecting that mileage decreases steeply for low top-speed values and more gradually for large top-speed values. The flat region in the middle of the plot is a feature that would probably be invisible on an examination of the raw data. This graph was produced using the “lowess” function (a local regression technique very similar to the loess function) of the free-source statistical software package R.

—David B. Hitchcock

See also Exploratory Data Analysis; Regression Analysis

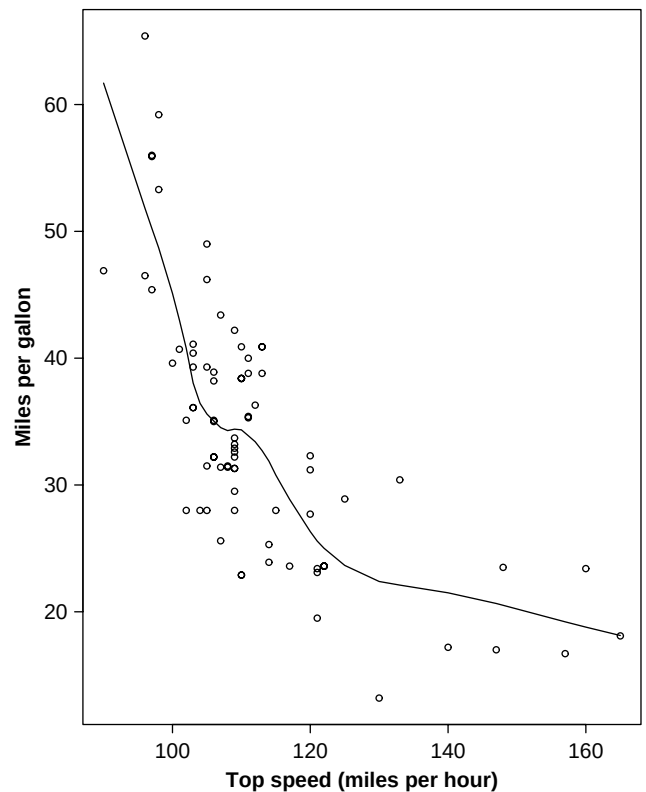


Figure 1 The Nonparametric Regression of Cars' Mileage on Their Top Speeds Yields This Smooth-Curve Estimate of the Relationship Between These Variables

Further Reading

- Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The Kernel Approach with S-Plus illustrations*. New York: Oxford University Press.
- Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Smoothing methods applet for fitting a regression curve to simulated data: <http://www.spss.com/research/wilkinson/Applets/smoothers.html>

SOCIAL CLIMATE SCALES

The Social Climate Scales comprise 10 separate scales; each taps the social climate of a different setting. The first group, the Family Environment Scale, Group Environment Scale, and Work Environment Scale,

measure the social climate in community settings. The Classroom Environment Scale and University Residence Environment Scale tap the social climate in educational settings. The Community-Oriented Programs Environment Scale, Ward Atmosphere Scale, and Sheltered Care Environment Scale focus on treatment and residential care settings, and the Correctional Institutions Environment Scale and Military Environment Inventory measure the social climate in these two settings. The Work Environment Scale is published by Consulting Psychologists Press (www.cpp.com), the Sheltered Care Environment Scale is available in a monograph published by Sage (www.sagepub.com), and the other scales are published by Mind Garden (www.mindgarden.com).

Each of the scales measures relationship, personal growth or goal orientation, and system maintenance and change dimensions. Relationship dimensions assess the quality of personal relationships in a setting. They measure how involved people are in a setting, how much they help each other, and how openly they express feelings.

Personal growth or goal orientation dimensions tap how an environment encourages specific goals and directions of change. For example, in the family, these dimensions reflect independence, achievement, intellectual and cultural orientation, participation in recreation, and moral and religious values. In the workplace, they assess employees' autonomy, task orientation, and work pressure.

System maintenance and change dimensions measure order and organization in a setting, how clear it is in its expectations, how much control it maintains, and how responsive it is to change.

Each of the Social Climate Scales has three forms: the Real Form (Form R), the Ideal Form (Form I), and the Expectations Form (Form E). The Real Form asks people how they see a social environment. The Ideal Form asks about a preferred or ideal setting. The Expectations Form asks people to describe what they imagine a new, unfamiliar environment will be like. These forms of the scales measure how people see the settings they are in, what an ideal setting for them would include, and what they expect a new setting they are about to enter will be like.

Each of the scales is based on extensive normative and psychometric data; the subscales have adequate internal consistency, stability, and construct and predictive validity. The scales have many practical applications. They are used to educate clinicians, consultants, and program evaluators about social environments and to help them describe environments, formulate clinical case descriptions, facilitate individual and family counseling, match individuals and environments, and plan interventions to improve environments. With respect to research applications, the scales are used to compare and contrast environments, understand the determinants and impacts of social climate, predict adjustment to life crises, and identify environmental risk factors.

—Rudolf H. Moos

See also Family Environment Scale

Further Reading

Moos, R. (2003). *The Social Climate Scales: A user's guide* (3rd ed.). Redwood City, CA: Mind Garden.

Moos, R., & Lemke, S. (1996). *Evaluating residential facilities*. Thousand Oaks, CA: Sage.

SOCIAL SKILLS RATING SYSTEM

The Social Skills Rating System (SSRS) provides a broad, multirater assessment of student social behaviors that can affect teacher-student relations, peer acceptance, and academic performance. The SSRS documents the perceived frequency and importance of behaviors influencing a student's development of social competence and adaptive functioning at school and at home. The SSRS components include three behavior rating forms (teacher, parent, and student versions) and an integrative assessment and intervention planning record. Teacher and parent forms are available for three developmental levels: preschool, Grades kindergarten through 6, and Grades 7 through 12. The SSRS assesses the domains of social skills, problem behavior, and academic competence.

The Social Skills Scale has five subscales: Cooperation, Assertion, Responsibility, Empathy, and Self-Control. The Problem Behaviors Scale has three subscales that measure Externalizing Problems, Internalizing Problems, and Hyperactivity. The Academic Competence domain concerns student academic functioning and consists of a small, yet critical, sample of relevant behaviors. Items in this domain are rated on a 5-point scale that corresponds to percentage clusters (1 = *lowest 10%*, 5 = *highest 10%*). This domain includes items measuring reading and mathematics performance, motivation, parental support, and general cognitive functioning. The scale appears on the teacher form at the elementary and secondary levels.

Scores are based on the results of 3-point ratings on the Social Skills Scale and the Problem Behaviors Scale (0 = *never occurs*, 1 = *sometimes occurs*, and 2 = *very often occurs*). All social skills items are rated on two dimensions: frequency and importance. The inclusion of the importance dimension allows raters to specify how important each social skill is for classroom success (teacher ratings), for their child's development (parent ratings), and for the student's relationships with others (student ratings).

The SSRS was standardized on a national sample of 4,170 children and used their self-ratings and ratings of them by 1,027 parents and 259 teachers. Internal consistency estimates for the SSRS across all forms and levels yielded a median coefficient alpha for the Social Skills Scale of .90, .84 for the Problem Behaviors Scale, and .95 for the Academic Competence Scale. Overall, these coefficients indicate a relatively high degree of scale homogeneity. The SSRS manual presents a number of studies investigating the construct, criterion-related, and content validity of the scale. For example, the SSRS Social Skills Scale and subscales correlate highly with other measures of social skills, such as the Walker-McConnell Scale of Social Competence and School Adjustment and the Social Behavior Assessment. The SSRS Problem Behaviors Scale and subscales show moderate to high correlations with the Child Behavior Checklist and the Harter Teacher Rating Scale. In addition, the SSRS Social Skills,

Problem Behaviors, and Academic Competence scales reliably differentiate children with mild disabilities (e.g., learning disabilities, behavior disorders, and mild mental or cognitive disabilities) from students without disabilities.

The SSRS has been used in hundreds of studies of children's social behavior. The SSRS has a representative national standardization and extensive evidence for reliability and validity. One of the most attractive features of the SSRS is its utility in selecting target behaviors for intervention purposes, a feature uncommon to most behavior rating scales. An intervention guide that provides a manualized approach to teaching and improving more than 40 social skills coordinates with the SSRS results.

—Stephen N. Elliott

Further Reading

- Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the child behavior checklist and revised child behavior profile*. Burlington: University of Vermont, Department of Psychiatry.
- Elliott, S. N., & Gresham, F. M. (1991). *Social skills intervention guide: Practical strategies for social skills training*. Circle Pines, MN: American Guidance Service.
- Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Service.
- Harter, S. (1985). *Manual for the self-perception profile for children*. Denver, CO: University of Denver.
- Stephens, T. (1978). *School skills in the classroom*. Columbus, OH: Cedars Press.
- Walker, H. M., & McConnell, S. (1988). *Walker-McConnell Scale of Social Competence and School Adjustment*. Austin, TX: PRO-ED.

SOCIETY FOR RESEARCH IN CHILD DEVELOPMENT

The Society for Research in Child Development (SRCD) is an organization of more than 6,000 members from various academic disciplines concerned with the study of child development. Disciplines include, but are not limited to, anthropology, home economics, linguistics, neuroscience, nursing, nutrition,

pediatrics, psychiatry, psychology, public health, and sociology. In addition to being multidisciplinary, the society's membership is also international.

The child development movement in the United States began in the 1920s, but its roots reach back much further. It arose from external pressures broadly based on desires for better health, rearing, education, legal rights, and occupational treatment of children. Movements related to child health, child study, and mental hygiene were prominent by the late 19th and early 20th centuries. From these activities came the idea that children were the responsibility of the society at large.

The scientific status of the field of child development received formal recognition in 1923 through the Committee on Child Development, appointed by the National Research Council, a division of the National Academies of Science. Its purpose was to integrate research activities and to stimulate research in child development. In 1927, *Child Development Abstracts and Bibliography*, SRCD's first publication (which continued until 2000), was begun.

In 1933, the Committee on Child Development formed a new, independent organization, and the Society for Research in Child Development emerged. Its mission is to promote multidisciplinary research on infant, child, and adolescent development in diverse contexts and its life-long implications; to foster the exchange of information among scientists and other professionals worldwide; and to encourage applications of research-based knowledge.

The society currently publishes three journals:

1. *Child Development*—contains original articles on development research and theory; six issues a year
2. *Monographs of the Society for Research in Child Development*—consists primarily of comprehensive reports of large-scale research projects or integrated programs of research; four issues a year
3. *Social Policy Report*—focuses each issue on a single topic affecting children, youth, or families and includes analyses of legislation and syntheses of research on issues of social policy and children; four issues a year

The society hosts a biennial meeting with an attendance of more than 5,500. These internationally attended meetings include individual research reports, symposia, invited lectures, and discussion sessions, among other timely and historical programs.

Almost one fifth of SRCD's members come from more than 50 nations other than the United States. Special efforts are made by the society to increase communication and interaction among researchers in human development throughout the world. The society also has a commitment to research and training in diversity. A major goal of SRCD is to increase and disseminate research on children from many racial and ethnic minorities.

The society maintains the Office of Policy and Communications in Washington, D.C. The director supervises the SRCD Fellows Program in Child Development. The goals of this program are to contribute to the effective use of scientific knowledge, to educate the scientific community about public policy, and to establish effective liaisons between scientists and federal and congressional offices. Fellows spend a year in federal agencies or congressional offices working to facilitate the application of research to societal issues regarding children and families. The Office of Policy and Communications also has programs to communicate research to the media and the public.

The society welcomes persons interested in child development. Membership is open to any individual actively engaged in research in human development or any of the related basic sciences, teaching relevant to human development, or activity otherwise furthering the purposes of the society. SRCD is located at 3131 S. State St. Suite 301, Ann Arbor, MI, 48108.

—John W. Hagen, Mary Ann McCabe,
and Nicholas G. Velissaris

Further Reading

Society for Research in Child Development Web site:
www.srkd.org

SOCIOLOGICAL ABSTRACTS

Sociological Abstracts, owned by Cambridge Scientific Abstracts (CSA), is an electronic database of bibliographic citations and abstracts that provides access to the world's literature in theoretical and applied sociology and related disciplines in the social and behavioral sciences. Fields covered range from anthropology to gender studies to substance abuse. Containing more than 650,000 records as of July 2005, the database provides citations and abstracts of articles from more than 1,800 journals from 1963 to the present. It also contains abstracts of books, book chapters, book reviews, conference papers, and dissertations, along with book, film, and software review citations. Hard-copy records published in the first 10 years of the database (1952–1963) were digitized and added to the electronic version of Sociological Abstracts in 2005. The database is updated on a monthly basis, with roughly 30,000 records added each year.

The journals presently reviewed for inclusion in the database are divided into three categories—core journals, priority journals, and selective sources—on the basis of the level-of-relevant-coverage rating they receive. Core journals are those produced by dedicated sociological associations, faculties, and so forth, or which include *sociology* in their titles. All significant articles appearing in core journals are fully abstracted and indexed. Priority journals, from disciplines such as anthropology and political science, which frequently address noteworthy sociological issues, are those that regularly feature the work of sociologists. More than half of the significant articles appearing in priority journals are covered in Sociological Abstracts. Serial periodicals that publish the works of sociologists intermittently are considered selective sources; less than half of the significant articles featured in these works are selected for coverage.

Sociological Abstracts contains both limited and comprehensive search tools. The Quick Search option allows users to enter a word or phrase pertaining to a subject of interest and to delimit the search by date.

The Advanced Search option permits even more elaborate search strategies through the use of various parameters. Database users may structure their searches through any number of combinations of fields, such as author, keyword, title, publisher, language, and so forth. They can further refine their searches by restricting their results to year or to journal articles only. In addition, Sociological Abstracts contains a thesaurus function to assist in expanding searches.

On completing a search, the database will display a list of resulting records, each showing the title, author, and source of the item, along with an excerpt from the item abstract. Sociological Abstracts does not provide access to full-text versions of the listed records. Users interested in obtaining sociology-related full-text documents should access an alternate CSA database, *Sociology: A SAGE Full-Text Collection*. CSA sells Sociological Abstracts database licenses to both institutions and individuals. Numerous academic libraries provide access to Sociological Abstracts to faculty, staff, and students.

—Johnny Holloway

Further Reading

Chall, M., & Owen, T. M. (1995). Documenting the world's sociological literature: Sociological Abstracts. *Publishing Research Quarterly*, 11(3), 83–95.

Sociological Abstracts: <http://www.csa.com/factsheets/socioabs-set-c.php>

SPATIAL LEARNING ABILITY TEST

The Spatial Learning Ability Test (SLAT) is a non-verbal measure of complex spatial manipulations (i.e., mental rotation and folding of a two-dimensional item into a three-dimensional item). A cognitive design system approach was applied to develop SLAT so that variations in items reflect variations in spatial processing. Both the level and the cognitive source of SLAT item difficulties are well predicted from a cognitive model that is operationalized by item stimulus

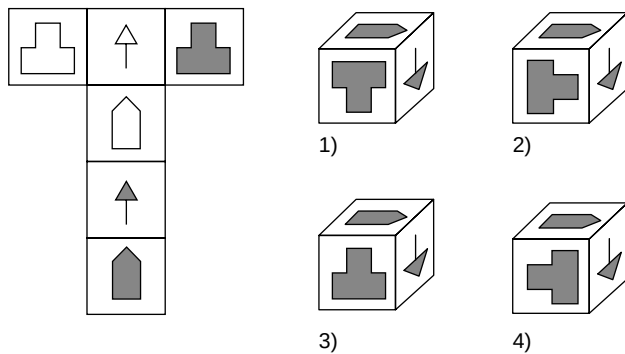


Figure 1 SLAT Cube-Folding Tasks

content. Research has supported SLAT as being a more pure measure of spatial processing than several other spatial tests, even when compared with tests with the same item type. The cognitive design system approach reduces the role of verbal-analytic process strategies in item solution, especially with the design of the distractors.

SLAT consists of cube-folding tasks, as shown in Figure 1. The stem of each SLAT item is a two-dimensional figure consisting of six connected squares with markings on each side. This two-dimensional figure can be folded into a cube. The alternatives show various three-dimensional views of folded cubes. The examinee must choose which alternative shows the correct representation of the markings on the sides of the cube when the cube has been folded.

Three fixed-content test forms and an adaptive form of SLAT are available for use as computerized tests. All SLAT items are calibrated by the Rasch item response theory model. The metric of SLAT scores is defined by the item response theory model. Norms for two samples of young adults are available. SLAT is also available in a dynamic form, which has some support for incremental validity over the static SLAT.

—Jennifer Ivie and Susan Embretson

Further Reading

- Embretson, S. E. (1992). *Technical manual for the Spatial Learning Ability Test* (Tech. Rep. No. 9201). Lawrence: University of Kansas, Department of Psychology.
- Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.),

Cognitive assessment: A multidisciplinary perspective. New York: Plenum.

Embretson, S. E. (1996). Cognitive design systems and the successful performer: A study on spatial ability. *Journal of Educational Measurement*, 33, 29–39.

Embretson, S. E. (1997). The factorial validity of scores from a cognitively designed test: The Spatial Learning Ability Test. *Educational and Psychological Measurement*, 57, 99–107.

Adaptive testing Web site including SLAT: <http://psychology.gatech.edu/cml/Adaptive/slat.htm>

Susan Embretson biography: <http://www.psychology.gatech.edu/WhoWeAre/Faculty/bio-SEmbretson.htm>

Spatial Statistics

The terms *spatial statistics* and *spatial analysis* refer broadly to that class of statistical methods used to analyze *spatial data*. Spatial data are observations for which the locations of the observations are an important part of the information carried by the data. It is common to divide the types of spatial data, and the associated analytical tools, into three groups: point pattern data, geostatistical data, and lattice data.

Point Pattern Data

For point pattern data, the locations of the observations constitute the attribute of interest. For example, such data might consist of the locations of all cases of childhood leukemia in a given state. Of interest is whether the data exhibit clustering. That is, do the locations of cases arise in groups more than would be expected from a random distribution of locations in the state? If so, this clustering might suggest an environmental cause associated with the disease.

In other instances, the pattern of locations might be regular. That is, there might be a greater distance between locations than would occur at random. This often occurs in cities, for example, with respect to the locations of elementary schools. Such regularity of the spatial distribution corresponds to policy decisions reflecting the inefficiency of grouping schools close together.

Combinations of patterns can also exist. At one scale, stores often occur in clusters, such as in shopping malls, while at another scale, there is regularity: Malls tend to be distributed across cities rather than tightly concentrated in one subregion of a city.

The statistical tools used to analyze point pattern data range in complexity. Simplest are those that involve dividing the study region into equal-sized subregions and performing a chi-square test to assess whether the counts of locations within the subregions differ significantly from what would be expected under a random (uniform) distribution of observations. If the region of interest is divided into a hierarchy of nested squares, say, the repeated application of such tests can be used to detect spatial patterns at various scales.

While straightforward, the aforementioned approaches require the division of the study region into arbitrary subregions, the choice of which is not without consequence. Thus, there is interest in methods that do not require such subdivision. One example is *Ripley's K function*, based on interpoint distances. In this method, one first calculates the geographic distance between each pair of points, and then the cumulative distribution function of these distances is determined. Finally, by means of Monte Carlo simulations, tests can be constructed to determine whether there is evidence of clustering, randomness, or regularity, and at what scale or scales these patterns occur.

Still more sophisticated methods exist for describing point pattern data. One class of these methods involves the construction of statistical models wherein the observations are assumed to be from a Poisson distribution whose intensity varies across the study area. The focus is then on characterizing the intensity function. This method is typically quite computationally intense, but it has attracted considerable interest recently in the study of spatial clustering of human diseases. More generally, point pattern methods find applications in a variety of biological, environmental, physical, and social science settings.

Geostatistical Data

Like point pattern data, geostatistical data include the locations of the observations, denoted here as

coordinates (x_i, y_i) , representing, for example, longitude and latitude. However, the focus with geostatistical data is less on the locations than on the measurements or observations recorded at the locations, denoted $w(x_i, y_i)$.

Geostatistical data arise often in environmental studies such as the following example. Especially in rural regions, there is interest in measuring the concentration of contaminants such as atrazine in well water. Such data are collected by assaying water sampled from numerous wells located on farms across the region. The concentrations, w , plus the locations of the observations, (x, y) , comprise the geostatistical data. A goal in analyzing such data might involve estimating the population mean concentration for the region and constructing a confidence interval for that mean or testing a hypothesis concerning the mean.

The primary concern for geostatistical data is that the observations w might not be statistically independent. Consequently, the usual statistical tools, t tests, F tests, analysis of variance (ANOVA), and so forth, which rely on such independence, are not applicable in their standard form. Instead, new methods must be used that take into account spatial autocorrelation.

An approach for describing spatial autocorrelation is based on several assumptions, including that of *anisotropy*. Put simply, anisotropy requires the correlation between two w values to depend only on the distance between observations and not on the direction from one to the other. Granting such an assumption, consider the value of w at two locations, (x_1, y_1) and (x_2, y_2) , a distance $h = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ apart. The *variogram* for observations separated by this distance is defined to be the expected value of the squared difference in w values: $E[w(x_1, y_1) - w(x_2, y_2)]^2$. With appropriate assumptions, it can be shown that this equals $2\sigma^2[1 - \rho(w(x_1, y_1), w(x_2, y_2))]$, where σ^2 represents the variance of a single observation, and ρ represents correlation. The variogram, therefore, can be viewed as a measure of correlation in spatial observations.

It is common to estimate the variogram as a function of the distance h . With such an estimate available, it is possible to adapt standard methods, eventually

yielding the spatial analogs of regression, ANOVA, confidence intervals, and so forth.

The majority of geostatistical tools apply best to Gaussian data. Other forms of continuous data might be approached by making an appropriate transformation of the observations, but the geostatistical analysis of categorical data, for example, is not straightforward, and research into such methods is ongoing.

Lattice Data

Lattice data bear similarities to geostatistical data. Again, greater interest is focused on some measured or observed quantity w , and again there is concern that these observations are spatially correlated, in which case standard statistical methods cannot be used. However, instead of focusing specifically on locations, as for geostatistical data, lattice data are distinguished by a *neighborhood structure*, which forms the basis for modeling spatial autocorrelation.

For example, consider the number of housing units for each county in a state. The numbers of housing units in two neighboring counties might be highly correlated, perhaps as a result of latitude (reflecting climate), amenities (proximity to a sea shore, proximity to a large urban area), and so forth. To formalize this spatial autocorrelation requires the definition of neighboring counties. For example, two counties might be deemed neighbors if they have an adjacent boundary. Alternatively, two (possibly nonadjacent) counties might be declared neighbors if their seats are not separated by more than 50 miles.

Consider the *neighborhood weights matrix* A whose ij th component is 1 if counties i and j are neighbors and 0 otherwise. Also, let W represent the vector of w values for all counties, and let 1 be a vector of 1s. We might postulate the model

$$W = \mu 1 + e,$$

where e represents the correlated error variation in the values of w about μ . One possible model for e takes the form

$$e = \rho A e + \varepsilon.$$

Here, the components of ε are assumed to be independent, Gaussian variables with variance τ^2 , say. Then, this hierarchical combination of models induces a spatial autocorrelation among the w values. Proceeding requires the estimation of μ , ρ , and τ^2 .

Numerous extensions of this model exist. The neighborhood matrix A need not consist only of 0 and 1 values; it is possible to incorporate noninteger and even negative values to reflect partial or inverse neighborhood associations. Alternative models for e can also be used, as can models for W , including regression terms. Moreover, while the resulting model for W is different, it is possible to apply lattice methods to categorical data. In such cases, inferences are usually based on Monte Carlo methods.

The great flexibility of lattice models makes them appealing tools, and they see considerable use in demography, economics, public health, and so on. This flexibility comes at a price, however, because it is not always clear how to define the neighborhood matrix A or how to choose among the many models for e . Perhaps for related reasons, the contrasting geostatistical methods tend to find greater favor in environmental, ecological, and other settings where the definition of a neighborhood structure is less obvious but where the assumption of autocorrelation being a function simply of distance is tenable.

Conclusion

As a subdiscipline, spatial statistics is enjoying considerable growth, fueled by several major changes in the research world. First, the ability to gather spatial data has greatly increased—the network of earth-orbiting satellites is a prime example, providing atmospheric, geographical, biological, and environmental data. Second, the advent of Geographical Information Systems increases our ability to archive and process such data. Both of these advances can be credited in large part to the existence of increased computing power. A third benefit of such power is the advance in statistical analysis tools, such as Monte Carlo simulation, that permit high quality, effective analyses.

Despite these advances, numerous open questions in spatial statistics remain, and the need for additional and more refined tools persists.

—Murray Clayton

Further Reading

- Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.
- Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*. London: Hodder Arnold.
- Haining, R. P. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge, UK: Cambridge University Press.
- Ripley, B. D. (2004). *Spatial statistics*. London: Wiley.

SPEARMAN'S RHO

Since its inception more than a century ago, the correlation coefficient has remained one of the most commonly used statistical indices in empirical research. A correlation coefficient provides an index of the relationship between two variables, or bivariate relationship. When both variables are quantitative in nature and either (a) they depart from normality or (b) they represent ordinal scales of measurement, then a researcher can use Spearman's rho (ρ), also known as the Spearman rank correlation coefficient. Charles Spearman introduced the concept of Spearman's rho in his seminal article in 1904, titled "The Proof and Measurement of Association Between Two Things," published in the *American Journal of Psychology*.

Technically speaking, Spearman's rho, a nonparametric measure of association, is appropriate whenever each variable is measured on at least an ordinal scale and both variables are taken from any continuous bivariate distribution. Moreover, Spearman's rho does not require any assumption about the frequency distribution of the variables. However, when the variables are normally distributed, the procedure known as Pearson's r , or the Pearson product-moment correlation, has more statistical power than does Spearman's rho and thus is more appropriate. In fact,

the asymptotic relative efficiency of Spearman's rho with respect to Pearson's r is 91.2%. In other words, Pearson's r is more efficient than Spearman's rho in that for a desired level of statistical significance (i.e., Type I error), Pearson's r has the same power for detecting statistical significance as does Spearman's rho using 91.2% of the sample size needed for Spearman's rho. Alternatively stated, if a sample size of 1,000 is needed for a relationship to be declared statistically significant for a nominal alpha level using Spearman's rho, then a smaller sample size (i.e., 912) is needed to yield the same p value when using Pearson's r . However, when the normality assumption is violated, the Type I error rate is inflated when Pearson's r is used, and thus nonparametric alternatives should be employed.

Spearman's rho is the second-most-popular bivariate correlational technique. Computation of Spearman's rho is as follows: Suppose we have n pairs of continuous data. The first step in computing Spearman's rho is to rank the X scores in the paired sample data from 1 (smallest score) to n (largest score), and independently rank the Y scores in the paired sample data from 1 (smallest score) to n (largest score). Therefore, the original $(X_1Y_1), (X_2Y_2), \dots, (X_nY_n)$ pairs of observations will change to $[\text{Rank}(X_1), \text{Rank}(Y_1)], [\text{Rank}(X_2), \text{Rank}(Y_2)], \dots, [\text{Rank}(X_i), \text{Rank}(Y_i)]$. The next step in the process is to calculate a difference d for each pair as the difference between the ranks of the corresponding X and Y variables. The sum of the differences always will equal zero. The test statistic, denoted by r_s , is defined as the sum of squares of these differences³. The formula for this expression is

$$r_s = 1 - \frac{6 \sum d_i^2}{(n^3 - n)}. \quad (1)$$

An examination of the extreme cases helps us see how this formula works. Suppose both X and Y scores represent identical ordered arrays from smallest to largest. In this case, each $d_i = 0$, and thus $\sum d_i^2 = 0$, and substitution in Equation 1 provides an r_s value of +1. Thus, $r_s = 1$ describes a perfect positive relationship between ranks. Conversely, if the ranks of the X scores

are the complete opposite of the ranks of the Y scores, then it can be demonstrated that

$$\sum d_i^2 = \frac{n(n^2 - 1)}{3} = \frac{(n^3 - n)}{3}. \quad (2)$$

Substitution of this value into Equation 1 leads to an r_s value of -1 , which describes a perfect negative relationship. Interestingly, Spearman's rho is equal to Pearson's r with ranks substituted for the original scores.

If two or more scores within the X or the Y set, or both, have the same value (i.e., tied observations), midranks are assigned to these observations. The midrank of a set of tied scores is defined as the average of the ranks to which they would have been assigned if they were not tied. Unfortunately, when one or more ties prevail, it is no longer possible for r_s to take the values -1 or $+1$ even when the association is perfect. As a result, we need to make an adjustment or correction for ties. The formula for r_s then becomes

$$r_s = \frac{n^3 - n - 6 \sum d_i^2 - 6(t' + u')}{\sqrt{n^3 - n - 12t'} \sqrt{n^3 - n - 12u'}}, \quad (3)$$

where $t' = (\sum t^3 - \sum t)/12$ for t , which represents the number of tied scores at any given rank in the X set, and the sums $\sum$ are over all the sets of t tied X ranks. Similarly, $u' = (\sum u^3 - \sum u)/12$ for ties in the Y set. Again, Spearman's rho in Equation 3 is equal to Pearson's r with ranks substituted for the original scores and with midranks used for ties. It should be noted that if there are no ties in either set, $t' = u' = 0$, then the formula for r_s in Equation 3 reduces to the expression in Equation 1.

Example With Real Data

Spearman's Rho as a Descriptive Statistic

In order to show how r can be computed, we will introduce some real data. Onwuegbuzie was interested in determining a relationship between the total number of points scored (x) by professional National Football League (NFL) teams and their winning

percentages (y) during the 1997–1998 football season. (Both variables are continuous.) All NFL teams were included in the study, yielding a sample size of 30. These data are presented in Table 1. Also, the scatterplot is displayed in Figure 1.

Although the NFL data represented scores that were normally distributed, with small skewness (.67 and .31 for total number of points scored and winning percentage, respectively) and kurtosis (.91 and $-.69$ for total number of points scored and winning percentage, respectively) coefficients, as can be seen from the SPSS output in Figure 2, we will use these data to demonstrate how to use Spearman's rho to assess the relationship between the total number of points scored (x) and winning percentage (y). The first step is to rank the X scores in the paired sample data from 1 (smallest score) to n (largest score), and independently rank the Y scores in the paired sample data from 1 (smallest score) to n (largest score). These ranks are displayed in Table 2.

Once the scores have been ranked and the difference in ranks calculated, squared, and summed, the next step is to compute the correction for ties. There are no ties in the number of points scored (i.e., x set), and so $t = 0$. However, for the winning percentage variable (i.e., y set), there are 11 sets of ties, as shown in Table 3. From Table 3, we have $u' = (281 - 29)/12 = 21.00$.

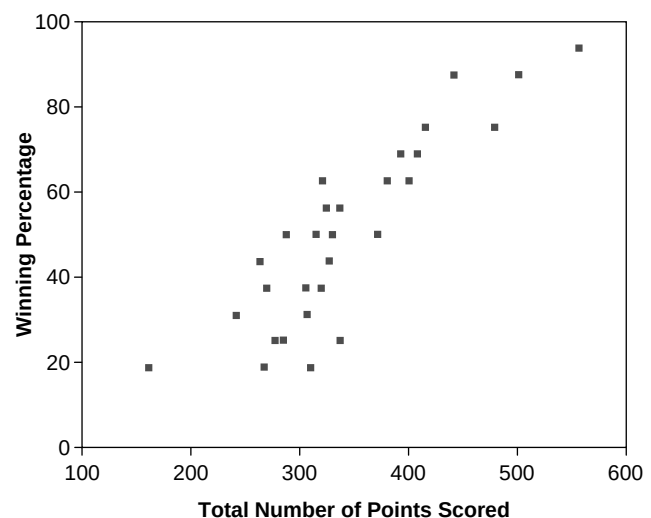


Figure 1 Scatter Plot of Number of Points Scored Versus Winning Percentage

Table 1 Spearman's Rho Analysis for National Football League Data, 1997–1998

<i>NFL Football Team</i>	<i>No. Points Scored (x)</i>	<i>Winning % (y)</i>	<i>Rank of x</i>	<i>Rank of y</i>	<i>d</i>	<i>d²</i>
New York Giants	287	50.00	8	16.0	−8.00	64.00
Washington Redskins	319	37.50	14	10.0	4.00	16.00
Philadelphia Eagles	161	18.75	1	2.0	−1.00	1.00
Dallas Cowboys	381	62.50	22	22.0	0.00	0.00
Arizona Cardinals	325	56.25	16	19.5	−3.50	12.25
Green Bay Packers	408	68.75	25	24.5	0.50	0.25
Tampa Bay Buccaneers	314	50.00	13	16.0	−3.00	9.00
Detroit Lions	306	31.25	11	7.5	3.50	12.25
Minnesota Vikings	556	93.75	30	30.0	.00	0.00
Chicago Bears	276	25.00	6	5.0	1.00	1.00
San Francisco 49ers	479	75.00	28	26.5	1.50	2.25
Carolina Panthers	336	25.00	19	5.0	14.00	196.00
Atlanta Falcons	442	87.50	27	28.5	−1.50	2.25
New Orleans Saints	305	37.50	10	10.0	0.00	0.00
St Louis Rams	285	25.00	7	5.0	2.00	4.00
New England Patriots	337	56.25	20	19.5	0.50	0.25
Miami Dolphins	321	62.50	15	22.0	−7.00	49.00
New York Jets	416	75.00	26	26.5	−0.50	0.25
Buffalo Bills	400	62.50	24	22.0	2.00	4.00
Indianapolis Colts	310	18.75	12	2.0	10.00	100.00
Pittsburgh Steelers	263	43.75	3	12.5	−9.50	90.25
Jacksonville Jaguars	392	68.75	23	24.5	−1.50	2.25
Tennessee Oilers	330	50.00	18	16.0	2.00	4.00
Cincinnati Bengals	268	18.75	4	2.0	2.00	4.00
Baltimore Ravens	269	37.50	5	10.0	−5.00	25.00
Kansas City Chiefs	327	43.75	17	12.5	4.50	20.25
Denver Broncos	501	87.50	29	28.5	0.50	0.25
Seattle Seahawks	372	50.00	21	16.0	5.00	25.00
Oakland Raiders	288	50.00	9	16.0	−7.00	49.00
San Diego Chargers	241	31.25	2	7.5	−5.50	30.25
Totals					0	724.00

Statistics			
		Winning %	Total Number of points scored
N	Valid	30	30
	Missing	0	0
Skewness		.309	.667
Std. Error of Skewness		.427	.427
Kurtosis		−.692	.906
Std. Error of Kurtosis		.833	.833

Figure 2 SPSS Output: Skewness and Kurtosis Coefficients for Variables in NFL Data Set

Substituting this information and that in Table 1 into Equation 3 gives

$$r_s = \frac{30^3 - 30 - 6(724.00) - 6(21.00)}{\sqrt{30^3 - 30 - 0}\sqrt{30^3 - 30 - 12(21.00)}} = .8382.$$

Spearman's Rho as an Inferential Statistic

Because Spearman's rho represents a descriptive measure of association, it can be used as a test statistic

Table 2 Spearman's Rho Analysis for National Football League Data

NFL Football Team	No. Points Scored (<i>x</i>)	Winning % (<i>y</i>)	<i>U</i>	<i>U</i> ²
Philadelphia Eagles	161	18.75	3	9
San Diego Chargers	241	31.25	3	9
Pittsburgh Steelers	263	43.75	2	4
Cincinnati Bengals	268	18.75	3	9
Baltimore Ravens	269	37.50	2	4
Chicago Bears	276	25.00	5	25
St Louis Rams	285	25.00	2	4
New York Giants	287	50.00	3	9
Oakland Raiders	288	50.00	2	4
New Orleans Saints	305	37.50	2	4
Detroit Lions	306	31.25	2	4
Indianapolis Colts	310	18.75		
Tampa Bay Buccaneers	314	50.00		
Washington Redskins	319	37.50		
Miami Dolphins	321	62.50		
Arizona Cardinals	325	56.25		
Kansas City Chiefs	327	43.75		
Tennessee Oilers	330	50.00		
Carolina Panthers	336	25.00		
New England Patriots	337	56.25		
Seattle Seahawks	372	50.00		
Dallas Cowboys	381	62.50		
Jacksonville Jaguars	392	68.75		
Buffalo Bills	400	62.50		
Green Bay Packers	408	68.75		
New York Jets	416	75.00		
Atlanta Falcons	442	87.50		
San Francisco 49ers	479	75.00		
Denver Broncos	501	87.50		
Minnesota Vikings	556	93.75		
Totals			29	85

Notes: *t* = the number of tied scores at any given rank in the *x* set; *U* = the number of tied scores at any given rank in the *y* set; *U*² = the square of the number of tied scores at any given rank in the *y* set. Because there are no ties in the *x* set, *t* = 0.

for the null hypothesis that *x* and *y* are not related (i.e., that *x* and *y* are independent) versus the alternative hypothesis that there is a positive relationship between the total number of points scored and the winning percentage. The sampling distribution of *r_s* is provided in Table 4, for *n* ≤ 30. In this table, the *p* values given are (a) in the right tail if the value of the test statistic (i.e., *r_s*) is positive, and (b) in the left tail if the value of

Table 3 Correction for Tied Data

Winning %	<i>U</i> (Frequency)	<i>U</i> ³
18.75	3	27
25.00	3	27
31.25	2	8
37.50	3	27
43.75	2	8
50.00	5	125
56.25	2	8
62.50	3	27
68.75	2	8
75.00	2	8
87.50	2	8
Total	29	281

the test statistic is negative. From Table 4, we can see that the right-tail *p* value is *p* < .001. Thus, we conclude that there is a positive relationship between the total number of points scored and the winning percentage.

For samples larger than 30, Spearman's rho that is uncorrected for ties is approximately normally distributed. Under the null hypothesis, *r_s* has a mean of zero and a standard deviation of

$$\frac{1}{\sqrt{n-1}}. \quad (4)$$

This implies that the *z* statistic for *r_s* is

$$Z = r_s \sqrt{n-1}, \quad (5)$$

where the *p* values are obtained from the standard normal tables. If there are ties in either sample, then the corrected *r_s* is used with Equation 5. However, when ties prevail, the exact null distribution typically cannot be provided because the *p* values depend on the particular configuration of ties. Therefore, the *p* values obtained from the corrected *r_s* are approximate. For our NFL data, we have

$$Z = .84\sqrt{30-1} = 4.52.$$

The approximate *p* value corresponding to this *z* statistic from the standard normal tables is *p* < .0002.

Table 4 Spearman's Rho Correlation Distribution

n	r_s	p	n	r_s	p	n	r_s	p	n	r_s	p
3	1.000	.167	7	.679	.055	8	.381	.180	9	.517	.081
	.500	.500		.643	.069		.357	.195		.500	.089
4	1.000	.042		.607	.083		.333	.214		.483	.097
	.800	.167		.571	.100		.310	.231		.467	.106
	.600	.208		.536	.118		.286	.250		.450	.115
	.400	.375		.500	.133		.262	.268		.433	.125
	.200	.458		.464	.151		.238	.291		.417	.135
	.000	.542		.429	.177		.214	.310		.400	.146
5	1.000	.008		.393	.198		.190	.332		.383	.156
	.900	.042		.357	.222		.167	.352		.367	.168
	.800	.067		.321	.249		.143	.376		.350	.179
	.700	.117		.286	.278		.119	.397		.333	.193
	.600	.175		.250	.297		.095	.420		.317	.205
	.500	.225		.214	.331		.071	.441		.300	.218
	.400	.258		.179	.357		.048	.467		.283	.231
	.300	.342		.143	.391		.024	.488		.267	.247
	.200	.392		.107	.420		.000	.512		.250	.260
	.100	.475		.071	.453	9	1.000	.000		.233	.276
	.000	.525		.036	.482		.983	.000		.217	.290
6	1.000	.001		.000	.518		.967	.000		.200	.307
	.943	.008	8	1.000	.000		.950	.000		.183	.322
	.886	.017		.976	.000		.933	.000		.167	.339
	.829	.029		.952	.001		.917	.001		.150	.354
	.771	.051		.929	.001		.900	.001		.133	.372
	.714	.068		.905	.002		.883	.002		.117	.388
	.657	.088		.881	.004		.867	.002		.100	.405
	.600	.121		.857	.005		.850	.003		.083	.422
	.543	.149		.833	.008		.833	.004		.067	.440
	.486	.178		.810	.011		.817	.005		.050	.456
	.429	.210		.786	.014		.800	.007		.033	.474
	.371	.249		.762	.018		.783	.009		.017	.491
	.314	.282		.738	.023		.767	.011		.000	.509
	.257	.329		.714	.029		.750	.013	10	1.000	.000
	.200	.357		.690	.035		.733	.016		.988	.000
	.143	.401		.667	.042		.717	.018		.976	.000
	.086	.460		.643	.048		.700	.022		.964	.000
	.029	.500		.619	.057		.683	.025		.952	.000
7	1.000	.000		.595	.066		.667	.029		.939	.000
	.964	.001		.571	.076		.650	.033		.927	.000
	.929	.003		.548	.085		.633	.038		.915	.000
	.893	.006		.524	.098		.617	.043		.903	.000
	.857	.012		.500	.108		.600	.048		.891	.001
	.821	.017		.476	.122		.583	.054		.879	.001
	.786	.024		.452	.134		.567	.060		.867	.001
	.750	.033		.429	.150		.550	.066		.855	.001
	.714	.044		.405	.163		.533	.074		.842	.002

(Continued)

(Continued)

n	r_s	p	n	r_s	p	n	r_s	p	n	r_s	p
10	.830	.002	10	.612	.067	10	.394	.235	10	.176	.316
	.818	.003		.600	.072		.382	.246		.164	.328
	.806	.004		.588	.077		.370	.257		.152	.341
	.794	.004		.576	.083		.358	.268		.139	.354
	.782	.005		.564	.089		.345	.280		.127	.367
	.770	.007		.552	.096		.333	.292		.115	.379
	.758	.008		.539	.102		.321	.304		.103	.393
	.745	.009		.527	.109		.309	.132		.091	.406
	.733	.010		.515	.116		.297	.139		.079	.419
	.721	.012		.503	.124		.285	.148		.067	.433
	.709	.013		.491	.033		.273	.156		.055	.446
	.697	.015		.479	.037		.261	.165		.042	.459
	.685	.017		.467	.040		.248	.174		.030	.473
	.673	.019		.455	.044		.236	.184		.018	.486
	.661	.022		.442	.048		.224	.193		.006	.500
	.648	.025		.430	.052		.212	.203			
	.636	.027		.418	.057		.200	.214			
	.624	.030		.406	.062		.188	.224			

Right-Tail (Left-Tail) Probability on r_s ($-r_s$) for One-Sided Test

11	.427	.536	.618	.709	.764	.855
12	.406	.503	.587	.678	.734	.825
13	.385	.484	.560	.648	.703	.797
14	.367	.464	.538	.626	.679	.771
15	.354	.446	.521	.604	.657	.750
16	.341	.429	.503	.585	.635	.729
17	.329	.414	.488	.566	.618	.711
18	.317	.401	.474	.550	.600	.692
19	.309	.391	.460	.535	.584	.675
20	.299	.380	.447	.522	.570	.660
21	.292	.370	.436	.509	.556	.647
22	.284	.361	.425	.497	.544	.633
23	.278	.353	.416	.486	.532	.620
24	.275	.344	.407	.476	.521	.608
25	.265	.337	.398	.466	.511	.597
26	.260	.331	.390	.457	.501	.586
27	.255	.324	.383	.449	.492	.576
28	.250	.318	.376	.441	.483	.567
29	.245	.312	.369	.433	.475	.557
30	.241	.307	.363	.426	.467	.548
	.200	.100	.050	.020	.010	.002

Tail probability on $|r_s|$ for two-sided test

For $n > 30$, the probabilities not found using the normal distribution table by calculating $Z = r_s \sqrt{n-1}$. The left- or right-tail probability for r_s can be approximated by the left- or right-tail probability for Z , respectively.

Source: Adapted from Gibbons (1993).

Note: Entries labeled under the p column represent the cumulative probability, right-tail from the value of r_s to its maximum value of 1, for all $r_s \geq 0$, $n \leq 10$. The same probability is a cumulative left-tail probability, ranging from -1 to $-r_s$. For $10 < n \leq 30$, the table yields the smallest value of r_s (largest value of $-r_s$) for which the right-tail (left-tail) probability for a one-sided test is less than or equal to selected values, .100, .050, .010, .005, .001, shown in the top row. The same values correspond to $|r_s|$ for a two-sided test with tail probability .200, .100, .020, .010, .002, shown on the bottom row.

Correlations

		WINPCT	TOTPTS
WINPCT	Pearson correlation	1.000	.866**
	Sig.(2-tailed)	.	.000
	N	30	30
TOTPTS	Pearson correlation	.866**	1.000
	Sig.(2-tailed)	.000	.
	N	30	30

Correlations

			WINPCT	TOTPTS
Spearman's rho	WINPCT	Correlation Coefficient	1.000	.838
		Sig. (2-tailed)	.	.000
		N	30	30
	TOTPTS	Correlation Coefficient	.838	1.000
		Sig. (2-tailed)	.000	.
		N	30	30

**. Correlation is significant at the .01 level (2-tailed)

Figure 3 SPSS Output for Pearson's r and Spearman's Rho

Comparison With Pearson's r

From the SPSS output in Figure 3, it can be seen that Pearson's r is .87. Thus, the Spearman's rho correlation of .84 represents 96.6% of the Pearson's r correlation, which reflects the loss in relative efficiency as a result of applying Spearman's rho on approximately normal data. Nevertheless, using Cohen's criteria of .1 for a small correlation, .3 for a moderate correlation, and .5 for a large correlation, Spearman's rho correlation still is extremely large.

—Anthony J. Onwuegbuzie,
Larry Daniel, and Nancy L. Leech

See also Kendall Rank Correlation

Further Reading

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Gibbons, J. D. (1993). *Nonparametric measures of association* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07B091). Newbury Park, CA: Sage.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.

Onwuegbuzie, A. J. (1999). Defense or offense? Which is the better predictor of success for professional football teams? *Perceptual and Motor Skills*, 89, 151–159.

Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools*, 9(1), 73–90.

Spearman's rho applet: http://faculty.vassar.edu/lowry/corr_rank.html (allows you to compute rho and perform a null hypothesis significance test pertaining to this rho value)

SPLIT HALF RELIABILITY

All observations and all measurements contain error. The focus of much work in measurement is to minimize and estimate the amount of error in any given measurement. In classical test theory, X is an observed score that is composed of

T , the true score, and E , the error score: $X = T + E$. The true score is never known, but it can be thought of as the long-range average of scores from a single instrument administered to an individual an infinite number of times (the expected value or expected score). The error score is random and may have many sources, including testing conditions, individual characteristics that fluctuate from administration to administration, differences in forms, or instability of an individual's ability or trait over time.

This random error score is quite different from systematic sources of error, such as testwiseness, which may systematically increase an individual's score on each administration. Since testwiseness is systematic or constant, it finds its way into the true score and creates problems regarding validity because the trait being measured inadvertently may be influenced by testwiseness. Random error, since it varies randomly, influences the consistency of scores but not the expected value of a score (the true score) and thus influences reliability, not validity.

Theoretically, we can estimate the amount of error if we know how much of a given score is due to errors of measurement. If we were able to test a single person

repeatedly without the effects of recall and fatigue, variation in their scores would be considered measurement error. If there were no measurement error, they would get the same score on each administration. Since it is not possible to test individuals repeatedly without the interference of recall and fatigue, we employ groups to estimate measurement error variance. This allows us to estimate the standard error of measurement, or the typical amount of measurement error in a set of scores.

If we take the classical test theory model of scores and consider groups of scores and their variances, we see that the variance of the observed scores equals the sum of the variance of true scores and the variance of error scores: $S_X^2 = S_T^2 + S_E^2$ (in sample notation).

This is the long way of introducing the need for reliability: Reliability is a tool used to estimate the standard error of measurement, but it also has some intrinsic benefits in and of itself. Theoretically, reliability is considered the correlation between scores on two parallel forms of a test. The idea is that if there is no measurement error at work, scores from two parallel forms administered to the same group of individuals should be perfectly correlated—each individual should obtain the same score. It can be shown that the correlation between two parallel forms of a test is equal to the ratio of true score variance to observed score variance, or the proportion of variance in observed scores that is due to true individual differences:

$$r_{tt} = \frac{S_T^2}{S_X^2}.$$

This reliability coefficient can then be used in estimation of the standard error of measurement because it tells us the proportion of observed variance that is true variance; the standard error of measurement is a function of the proportion of observed variance that is true variance.

Estimating Reliability

Based on the classical test theory conception of reliability, scores are reliable to the extent that individuals' scores remain constant on repeated measurements. One way to estimate the degree of consistency of scores is on the basis of the idea of internal consistency of

performance. One way to think about internal consistency is to split the test in half and consider each half a parallel form of the test, evaluating the consistency of performance on the two halves.

This form of reliability is best estimated when the actual measurement procedure employs a single form, in which error results from item or content sampling error. This is typically done by administering the single form to each individual and then selecting some method of splitting the test in half to obtain two scores for each individual. The two scores resulting from the two halves are then correlated. This correlation is the split half reliability coefficient.

Formal Conditions for Split Half Reliability

Split half reliability is estimated from the administration of a single form on a single occasion. This is appropriate when the measurement procedure includes only one form or the inference from scores is based on the ability to generalize beyond the items selected for that form. Knowledge of split half reliability allows one to generalize from one sample of items to a larger domain of items. To be a coefficient of reliability, the split halves should be equivalent in terms of content specifications and statistical properties (primarily item difficulties).

There are many methods of splitting a test into halves. One could consider a careful systematic split securing equivalent halves in terms of content and difficulty. A more common approach is to divide items alternately, placing the odd-numbered items into one half test and the even-numbered items into the other half test. Since most tests are organized by content areas and difficulty, this approach provides essentially equivalent halves. Another method would be to simply randomly assign items to each half; however, random sampling could easily lead to nonequivalent halves. Whatever method is used, the correlation of scores from the two halves provides the split half reliability estimate.

Because this method of reliability estimation is based on a single form of a test and the inference regarding score consistency includes generalizing

from performance on a sample of items to performance on the domain of items, sampling error will be impacted by the number of items. All else equal, more items provides more consistent scores. Since we know that test length affects reliability, we need to adjust for the fact that we have a correlation from a test of half the length of the test of interest (the complete form). This is done by employing the Spearman-Brown prophecy formula. An estimate of the complete-form reliability is obtained by taking 2 times the split half reliability divided by 1 plus the split half reliability:

$$\hat{r}_{tt} = \frac{2r_{AB}}{1 + r_{AB}}.$$

This form of reliability recognizes that reliability can be estimated from consistent performance within a single test, based on two parallel or equivalent subtests. Note that the use of any one method to split the test in half is at least partially arbitrary—there is no theoretical justification for any given split. If the two halves are not equivalent, the resulting correlation will underestimate the reliability of the full test.

General Issues Regarding Reliability

Because the split half estimate of reliability is based on a correlation, it is not simply a characteristic of the measurement instrument itself. Score variability directly influences correlations, such that all else equal, the more score variance present, the higher the correlation and thus the higher the reliability. Correlational forms of reliability are sample specific and thus not necessarily generalizable to other samples. They do, however, provide an estimate of score consistency for the scores at hand.

In any estimate of reliability, conditions present during the specific administration of the measurement instrument can impact performance and scores in random ways, leading to lower consistency of scores and lower reliability. Each type of reliability estimate (e.g., split half reliability) also captures a specific form of random error. The split half reliability primarily captures measurement error due to sampling items from a domain, a form of sampling error. If this source of error is important to estimate given the measurement

procedure (because a single form is used, and we assume it is a random, representative sample of the domain), then it is an appropriate form of reliability. Technically speaking, an estimate of reliability should be obtained for each set of scores because any one estimate is sample specific, and the argument of generalizability across samples is difficult to make.

Finally, because sampling error is a function of sample size, all else equal, longer forms will yield higher reliability coefficients. Better, larger samples of items from the domain will reduce the likelihood that two forms differ in their ability to cover the domain. Also, because of the arbitrary aspect of methods to split a test form in half, other internal consistency measures are generally thought to be better, including coefficient alpha.

—Michael C. Rodriguez

See also Coefficient Alpha; Parallel Forms Reliability; Reliability Theory; Standard Error of Measurement

Further Reading

- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, Macmillan.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

SPREADSHEET FUNCTIONS

A spreadsheet function is a predefined formula. Excel, the most popular spreadsheet, has several categories of functions, including one labeled *statistical*.

One of the most simple of these functions is AVERAGE, which computes the average of a set of values. For example, the following statement averages the numbers in cells A1 through A3:

=AVERAGE(A1:A3)

The name of the function is AVERAGE, and the argument is A1:A3n.

A similar common function produces the sum of a set of cells:

`=SUM(A1:A3)`

In both cases, the results of these calculations are placed in the cell that contains the statement of the function. For example, to use the SUM (or any other) function, follow these steps:

1. Enter the function in the cell where you want the results to appear.

2. Enter the range of cells you want the function to operate on.
3. Press the Enter key, and voila! There you have it. Figure 1 shows the function, the argument, and the result.

Functions can be entered directly when the name of the function and its syntax are known or by means of the Insert command. Some selected Excel functions that perform statistical operations are shown in Table 1.

—Neil J. Salkind

Table 1 Excel Functions and What They Do

<i>Function Name</i>	<i>What It Does</i>
AVERAGE	Returns the average of its arguments
CHIDIST	Returns the one-tailed probability of the chi-square distribution
CHITEST	Returns the test for independence
CORREL	Returns the correlation coefficient between two data sets
FDIST	Returns the <i>F</i> probability distribution
FORECAST	Returns a value along a linear trend
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an <i>F</i> test
GEOMEAN	Returns the geometric mean
KURT	Returns the kurtosis of a data set
LINEST	Returns the parameters of a linear trend
MEDIAN	Returns the median of the given numbers
MODE	Returns the most common value in a data set
NORMDIST	Returns the normal cumulative distribution
NORMSDIST	Returns the standard normal cumulative distribution
PEARSON	Returns the Pearson product-moment correlation coefficient
QUARTILE	Returns the quartile of a data set
SKEW	Returns the skewness of a distribution
SLOPE	Returns the slope of the linear regression line
STANDARDIZE	Returns a normalized value
STDEV	Estimates standard deviation based on a sample
STDEVA	Estimates standard deviation based on a sample, including numbers, text, and logical values
STDEVP	Calculates standard deviation based on the entire population
STDEVPA	Calculates standard deviation based on the entire population, including numbers, text, and logical values
STEYX	Returns the standard error of the predicted y-value for each x in the regression
TDIST	Returns the Student <i>t</i> distribution
TREND	Returns values along a linear trend
TTEST	Returns the probability associated with a Student <i>t</i> test
VAR	Estimates variance based on a sample
VARA	Estimates variance based on a sample, including numbers, text, and logical values
VARP	Calculates variance based on the entire population
VARPA	Calculates variance based on the entire population, including numbers, text, and logical values

A11		fx =SUM(A1:A10)		
	A	B	C	D
1	5			
2	6			
3	5			
4	6			
5	7			
6	8			
7	7			
8	6			
9	5			
10	6			
11	61			
12				

Figure 1 Using the SUM Function as an Example

Further Reading

Spreadsheet functions usage instruction: <http://spreadsheets.about.com/od/excelfunctions/>

SPURIOUS CORRELATION

In social science research, the idea of spurious correlation is taken to mean roughly that when two variables correlate, it is not because one is a direct cause of the other but rather because they are brought about by a third variable. This situation presents a major interpretative challenge to social science researchers, a challenge that is heightened by the difficulty of disentangling the various concepts associated with the idea of spurious correlation.

Correlation and Causation

Drawing appropriate causal inferences from correlational data is difficult and fraught with pitfalls. One basic lesson social scientists learn in their undergraduate statistics education is that correlation does not imply causation. This adage is generally taken to mean that correlation *alone* does not imply causation. A correlation between two variables *X* and *Y* is not sufficient for inferring the particular causal relationship “*X* causes *Y*” because a number of alternative causal interpretations must first be ruled out. For example, *Y* may be the cause of *X*, or *X* and *Y* may be

produced by a third variable, *Z*, or perhaps *X* and a third variable, *Z*, jointly produce *Y*, and so on.

The statistical practice in the social sciences that is designed to facilitate causal inferences is governed by a popular theory of causation known as the *regularity theory*. This theory maintains that a causal relation is a regularity between different events. More specifically, a relationship between two variables *X* and *Y* can properly count as causal only when three conditions obtain: (a) *X* precedes *Y* in time; (b) *X* and *Y* covary; and (c) no additional factors enter into, and confound, the *X*-*Y* relationship.

The third condition requires a check for what social scientists have come to call *nonspuriousness*. A relationship between *X* and *Y* is said to be nonspurious when *X* is a direct cause of *Y* (or *Y* is a direct cause of *X*). A relationship between *X* and *Y* is judged nonspurious when we have grounds for thinking that no third variable, *Z*, enters into and confounds the *X*-*Y* relationship. In this regard, researchers typically seek to establish that there is neither a common cause of *X* and *Y* nor a cause intervening between *X* and *Y*.

Senses of Spurious Correlation

The term *spurious correlation* is ambiguous in the methodological literature. It was introduced by Karl Pearson at the end of the 19th century to describe the situation in which a correlation is found to exist between two ratios or indices even though the original values are random observations on uncorrelated variables. Although this initial sense of a spurious correlation remains a live issue for some social science researchers, it has given way to a quite different sense of spurious correlation. In the 1950s, Herbert Simon redeployed the term to refer to a situation where, in a system of three variables, the existence of a misleading correlation between two variables is produced through the operation of the third causal variable. H. M. Blalock’s extension of Simon’s idea into a testing procedure for more-complex multivariate models has seen this sense of a spurious correlation come to dominate in the social sciences. As a consequence, the social sciences have taken the problem of spuriousness to be equivalent to checking for the existence of third variables.

A Typology of Correlations

In order to understand that this problem of the third variable is not really a matter of spuriousness, it is important to be able to identify different types of correlations in terms of their presumed causes. The following is a typology of such correlations identified in terms of different kinds of presumed causes. These various correlations are sometimes confused when considering the problem of spuriousness.

At the most general level, the typology identifies two kinds of correlation: accidental and genuine. Accidental correlations are those that cannot be given a proper causal interpretation. There are two types of accidental correlation: nonsense and spurious. By contrast, genuine correlations are amenable to a proper causal interpretation. There are two types of genuine correlation: direct and indirect.

Nonsense correlations are those accidental correlations for which no sensible, natural causal interpretations can be provided. Statisticians delight in recounting the more amusing of these cosmic coincidences, such as the high positive correlation between birth rate and number of storks for a period in Britain or the negative correlation between birth rate and road fatalities in Europe over a number of years. In the statistics literature, these are sometimes called *illusory correlations*. These correlations exist, of course, but they cannot be given a plausible causal interpretation.

As characterized here, spurious correlations are accidental correlations that are not brought about by their claimed natural causes. To be true to their name, spurious correlations cannot be genuine correlations because they are false. They are artifacts of method and arise from factors such as sample selection bias; use of an inappropriate correlation coefficient; large sample size; or errors of sampling, measurement, and computation. Karl Pearson's original sense of spurious correlation mentioned above belongs here because the misleading value of ratio correlations depends, not on the relationship between the variables in question, but on their sharing of highly correlated components.

Direct correlations are genuine correlations for which one of the correlates is said to be a direct cause of the other. For example, heavy trucks are a direct

cause of road damage, and frequent and intense sun spots directly cause radio transmission noise. The social sciences are replete with empirical studies that are concerned with establishing direct causal relations. For example, manifest independent variables are examined in outcome studies on the assumption that they impact measured dependent variables in a causally direct way. Just as indirect correlations are often misleadingly called spurious correlations, so direct correlations are sometimes misleadingly called nonspurious correlations.

Indirect correlations are the genuine correlations that are produced by common or intervening causes and that we misleadingly call spurious correlations. However, there is nothing spurious about them at all. So-called spurious correlations are really genuine correlations, so their existence can hardly be denied by claiming that they are brought about by some underlying third variable. For example, if general intelligence is the common cause of correlated IQ performance on the verbal and numerical subtests of an intelligence test, then those subtest performances are indirectly and genuinely correlated. Clearly, this correlated IQ performance is not spurious because general intelligence explains why the correlation obtains; it does not render the correlation nonexistent or give us grounds for thinking that this is so.

From this analysis, it is clear that spurious correlations, properly named, must be regarded as a class of accidental correlations; otherwise we cannot sensibly deny the causal relations they are mistakenly thought to express.

Generative Causation

Tests for so-called spurious correlations are generally carried out to determine whether causal relations are empirical regularities. For this task, the regularity theory of causation is adequate, but only up to a point. Its requirements of temporal priority and covariation capture the idea of regular succession, but in order to properly understand the so-called problem of spuriousness, it is necessary to go beyond the restrictions of the regularity theory. A theory that allows us to do this is the generative theory of causation. The generative

theory depicts causation as a relation where, under appropriate conditions, a causal mechanism *produces* its effect. For this to happen, the causal mechanism must connect to its effect and have the power to generate that effect, usually when stimulated by the appropriate causal condition. It should be noted that it is the productivity of a generative mechanism that makes it a causal mechanism, and for this to be possible, there must be a naturally necessary connection that allows for a transmission of power from a cause to its effect. This causal power exists irrespective of whether it is currently being exercised. As such, it is properly viewed as a *tendency*, that is, an existing state of an object which, if unimpeded, will produce its effect. When it is unimpeded, we are able to diagnose the presence of the causal mechanism on the basis of the triggering condition(s) or its presumed effect(s) or both.

Unlike the regularity theory of causation, the generative theory is able to accommodate explanatory theories that are concerned with illuminating unobserved causal mechanisms. We need a theory of causation that affords us the conceptual space to do this because many of the world's causal mechanisms are not open to direct inspection. This is certainly true of the social sciences, where many of the postulated causal mechanisms are internal states. Intellectual abilities, personality traits, and emotional states are obvious cases in point.

Causation and Spuriousness

Simon's influential analysis of spurious correlation reveals a commitment to something like the regularity theory of causation. He notes that in order to distinguish true from spurious correlation, the term *cause* must be defined in an empiricist manner, with no reference to necessary connections between events, as the generative theory of causation makes.

Simon believes that a commitment to empiricist thinking about causality enables him to distinguish true from spurious correlations as he understands them. Ironically, however, this commitment actually prevents him from drawing his distinction properly. Recall that, for Simon, correlations are spurious if

they are brought about by common or intervening causes. Now, given that many of these causes will be latent or unobserved, it follows from a commitment to the regularity theory of causation that for a methodologically acceptable treatment of these variables to be possible, Simon and fellow empiricists must focus on altogether different variables at the manifest level. But this cavalier ontological attitude threatens to wreck our efforts to obtain genuine causal knowledge because the manifest replacement variables cannot act as surrogates for their latent variables, which are common and intervening causes. They are ontologically distinct from such causes and, although as causal conditions they may trigger their latent counterparts, they do not function as major causal mechanisms that determine so-called spurious correlations. Clearly, a coherent perspective on third variables that are latent variables requires a generative theory of causation.

Conclusion

When addressing the third variable problem, methodologists and researchers employ the misleading term *spurious correlation* to speak about genuine, indirect correlations. This practice only muddies the waters. Drawing causal inferences from correlational information is as difficult as it is important, and being clear about our key concepts can only facilitate such an undertaking. Not only is the terminology confusing, and thereby an impediment to understanding, but it also encourages a misleading view about the relation between causation and spuriousness that has the potential to misguide our causal modeling practices.

—Brian Haig

Further Reading

- Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10, 364–376.
- Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, 45, 1289–1298.
- Blalock, H. M. (1964). *Causal inference in nonexperimental research*. Chapel Hill: University of North Carolina Press.
- Haig, B. D. (2003). What is a spurious correlation? *Understanding Statistics*, 2, 125–132.

- Harré, R., & Madden, E. H. (1975). *Causal powers*. Oxford, UK: Blackwell.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60, 489–498.
- Prather, J. E. (1988). Spurious correlation. *Encyclopedia of statistical science* (Vol. 8, pp. 613–614). New York: Wiley.
- Simon, H. (1985). Spurious correlation: A causal interpretation. In H. M. Blalock (Ed.), *Causal models in the social sciences* (2nd ed., pp. 7–21). New York: Aldine.

STANDARD DEVIATION

The standard deviation (abbreviated as *s* or *SD*) represents the average amount of variability in a set of scores as the average distance from the mean. The larger the standard deviation, the larger the average distance each data point is from the mean of the distribution.

The formula for computing the standard deviation is as follows:

$$s = \sqrt{\sum \frac{(X - \bar{X})^2}{n - 1}},$$

where

s is the standard deviation,

Σ is sigma, which tells you to find the sum of what follows,

X is each individual score,

$\bar{X}$ is the mean of all the scores, and

n is the sample size.

This formula finds the difference between each individual score and the mean $X - \bar{X}$, squares each difference, and sums them all together. Then it divides the sum by the size of the sample (minus 1) and takes the square root of the result. As is apparent, the standard deviation is an average deviation from the mean.

Here is a sample data set for the manual computation of the standard deviation:

5, 4, 6, 7, 8, 6, 5, 7, 9, 5

Table 1 Computing the Standard Deviation

<i>X</i>	$(X - \bar{X})$	$(X - \bar{X})^2$
5	−1.2	1.44
4	−2.2	4.84
6	−.2	.04
7	.8	.64
8	1.8	3.24
6	−.2	.04
5	−1.2	1.44
7	.8	.64
9	2.8	7.84
5	−1.2	1.44
Sum	0	21.6

To compute the standard deviation, follow these steps:

1. List each score. It doesn't matter whether the scores are in any particular order.
2. Compute the mean of the group.
3. Subtract the mean from each score.
4. Square each individual difference. The result is the column marked $(X - \bar{X})^2$ in Table 1.
5. Sum all the squared deviations about the mean. As you can see in Table 1, the total is 21.6.
6. Divide the sum by $n - 1$, or $10 - 1 = 9$, so then $21.6/9 = 2.40$.
7. Compute the square root of 2.4, which is 1.55. That is the standard deviation for this set of 10 scores.

What we now know from these results is that each score in this distribution differs from the mean by an average of 1.55 points.

The deviations about the mean are squared to eliminate the negative signs. The square root of the entire value is taken to return the computation to the original units.

Every statistical package available computes the standard deviation. In Figure 1, Excel's Data Analysis ToolPak was used to compute a set of descriptive statistics, including the standard deviation.

	A	B	C	D
1	5		Column1	
2	4			
3	6		Mean	6.20
4	7		Standard Error	0.49
5	8		Median	6.00
6	6		Mode	5.00
7	5		Standard Deviation	1.55
8	7		Sample Variance	2.40
9	9		Kurtosis	-0.46
10	5		Skewness	0.48
11			Range	5.00
12			Minimum	4.00
13			Maximum	9.00
14			Sum	62.00
15			Count	10
16				

Figure 1 Using the Excel Data Analysis ToolPak to Compute the Standard Deviation

Summary

- The standard deviation is computed as the average distance from the mean.
- The larger the standard deviation, the more spread out the values are, and the more different they are from one another.
- Just like the mean, the standard deviation is sensitive to extreme scores.
- If $s = 0$, there is absolutely no variability in the set of scores, and they are essentially identical in value. This will rarely happen.

—Neil J. Salkind

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

Standard deviation computation steps and tutorial: http://davidmlane.com/hyperstat/desc_univ.html

STANDARD ERROR OF THE MEAN

The standard error of the mean (or *standard error*, for short) is a measure of how representative a sample is of the population from which it was drawn. Using information about how people vary within a sample and how big that sample is, we can estimate how

much variation we would expect to see if we drew multiple samples from a given population.

The concept of the standard error of the mean is similar to the standard deviation. Just as the standard deviation allows us to see how much individuals vary within a sample, so the standard error allows us to estimate how much samples will vary within a population. This is important because in research, it is seldom possible to observe the whole population we are interested in. We nearly always observe a smaller sample and make inferences from it to the population. For example, if we were interested in finding out whether Americans, on average, have a higher body mass index (BMI) than Europeans do, then we would take a sample of Americans and a sample of Europeans and find out the average BMI of both samples. It would be unfeasible to get this information for all members of both populations, but using the information we get from our sample, we can build up a picture about what the distribution of all possible samples would look like.

Take a look at Figure 1a. This shows a normal distribution of BMI scores in a sample of 50 Americans using some hypothetical data. The mean BMI score for this sample is 24.35. Based on the nature of normal distributions, the standard deviation of 3.31 tells us that 68% of all individuals in this sample have a

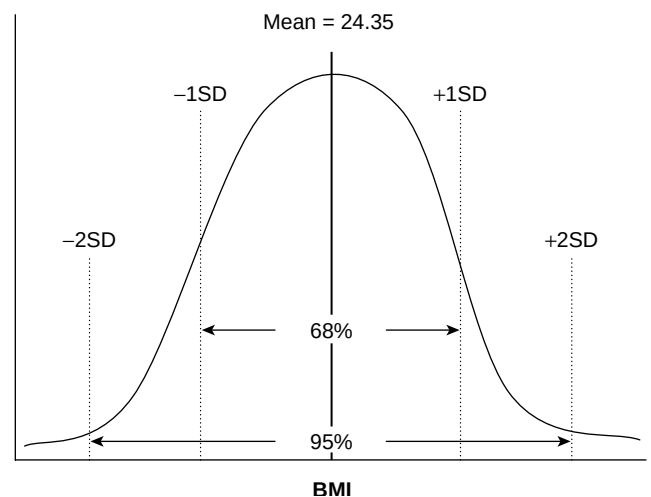


Figure 1a A Normal Distribution of Body Mass Index Scores From a Hypothetical Sample of 50 Americans

score between 21.04 and 27.66 (± 1 standard deviation) and that 95% have a score between 17.73 and 30.97 (± 2 standard deviations).

Now take a look at Figure 1b. This also shows a normal distribution, but this time it is for all possible sample means from the population of Americans (again using hypothetical data). Undoubtedly there will be some variation between the different possible samples we could have selected. We may have selected a particularly overweight group of Americans, for instance. Imagine we *could* take all possible samples. Each one would have a different makeup of people and thus would have a slightly different average BMI. The standard error tells us the degree to which sample means drawn from the same population are likely to differ.

Notice that the mean score of samples is the same as the mean score of the sample of 50 individuals. Notice also that the distribution of all sample means is narrower. This makes sense when we realize that we are less likely to observe extreme means than we are to observe extreme individuals. What is really important, though, is that we are now able to estimate the upper and lower boundaries of the mean that we expect to see if the sample has been drawn from this population. Using the standard error of .47,

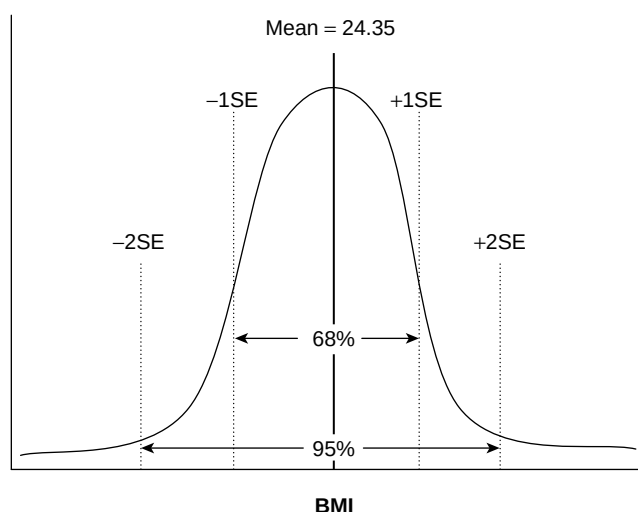


Figure 1b A Normal Distribution of Hypothetical Mean Body Mass Index Sample Scores From an American Population

we see that 68% of all sample means will be between 23.88 and 24.82, and 95% will be between 23.41 and 25.29. If we find a sample that has a mean score outside those limits, then we can assume that it is not from the same population (using 95% confidence intervals).

The concept of the standard error of the mean was first introduced by Ronald Fisher in 1925 in his book *Statistical Methods for Research Workers*. It is a concept that really underpins the frequentist philosophy of statistics that dominates the social sciences. Because if we can estimate the distribution of all possible sample means, then we can also estimate the probability of observing the mean we have observed in a given population and thus test hypotheses.

Assumptions of the Standard Error of Mean

1. The data are normally distributed.
2. The data occur at an interval or ratio level.

An Example

Let's take the example given earlier regarding BMI and nationality. Let's say we took 50 Americans and 50 Europeans and took their height and weight and then calculated a BMI for each person. The means and standard deviations appear in Table 1.

We find that the average BMI is 24.35 for the Americans and 20.69 for the Europeans in our study. Now let's calculate the standard error associated with each of these means. To do this, we need to know how wide the variability is between the scores in each

Table 1 A Comparison of Hypothetical Average Body Mass Index Scores for Americans and Europeans

Nationality	N	Mean	SD
American	50	24.35	3.31
European	50	20.69	1.47
Total	100	22.52	3.14

sample (the standard deviation, or σ) and how large the samples are (N).

$$\text{Standard error} = \sigma/\sqrt{N}.$$

You can see that the standard error will increase as the variability in the data (σ) increases and will decrease as the sample size (N) increases. So for Americans,

$$\text{standard error} = 3.31/7.07 = .47.$$

And for Europeans,

$$\text{standard error} = 1.47/7.07 = .21.$$

According to the principles underlying normal distributions, the standard error shows us that we can be 95% confident that the population BMI score for Americans is somewhere between 23.41 (mean – 2 standard errors) and 25.29 (mean + 2 standard errors) and that for Europeans, the population mean is somewhere between 20.27 and 21.11. Since the boundaries between the two groups do not cross, we can be 95% confident that the two samples are from different populations.

Calculating Standard Error Using SPSS

SPSS will calculate the standard error of a mean through the DESCRIPTIVES option or through the COMPARE MEANS option. As we have a grouping variable in this example, we will go through the COMPARE MEANS option.

1. Go to ANALYZE, COMPARE MEANS, and then MEANS. A command box will appear with all your variables on the left-hand side.
2. Select the variable of interest and arrow it into the *Dependent list* box.
3. Select the grouping variable and arrow it into the *Independent list* box.
4. Click on OPTIONS, find Std. Error of Mean in the left-hand box, and arrow it into the right-hand box.
5. Now click CONTINUE and OK. In the output, you will get a table that shows the standard error, along with the other descriptive statistics (see Table 2).

Table 2 The Standard Error and Other Descriptive Statistics

Report				
BMI				
Nationality	Mean	N	Std. Deviation	Std. Error of Mean
American	24.3520	50	3.31073	.46821
European	20.6900	50	1.47250	.20824
Total	22.5210	100	3.14400	.31440

If you wanted to look at the standard errors associated with each mean graphically, you could select the ERROR-BAR option in GRAPHS. For this example, we want to create a simple error bar graph for summaries of separate variables. Highlight the variables of interest and put them into the ERROR-BAR box. Figure 2 shows the resulting bar graph.

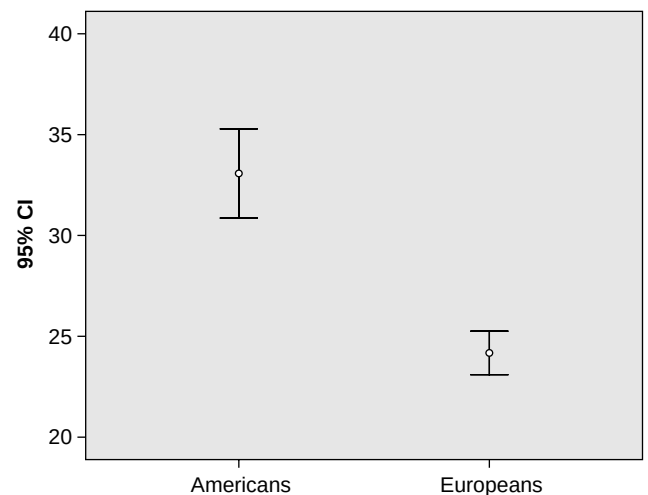


Figure 2 An Error Bar Graph Comparing Hypothetical Mean BMI Scores of Europeans and Americans

—Siân E. Williams

Further Reading

Fisher, R. A. (1925). *Statistical methods for research workers* (1st–13th eds.). Edinburgh, UK: Oliver & Boyd.

STANDARD ERROR OF MEASUREMENT

In order for the scores produced by a measure (referred to hereinafter as scores) to prove useful for the purpose of basic or applied research, it is critical that the measure be reliable. Reliability can be viewed from the perspective of systematic versus nonsystematic sources of variability in scores. To the degree that scores are systematic, they lead to measurement that is precise. Thus, if the level of an attribute (e.g., verbal ability) of a measured entity (e.g., person) remains unchanged, then repeated measurement of the attribute should produce scores that do not vary from one measurement occasion to the next. The greater the degree to which the variability in scores is a function of systematic variance, the more reliable the measure.

Scores are nonsystematic to the extent that they contain error variance that is random in nature. The greater the degree of error variance, the more scores will vary from one measurement occasion to the next. Error in the scores will cause them to vary across occasions even though the level of the measured attribute remains constant.

Ways of Viewing the Reliability of a Measure

There are two basic ways of characterizing the reliability of a measure. One is the *reliability coefficient* (ρ_{xx}), which varies between 0 (totally unreliable) and 1 (totally reliable).

Conceptually, it is the proportion of variance in scores that is systematic. It can be estimated through three general strategies. The *test-retest strategy* requires that a number (N) of entities be measured on two occasions with the same k -item measure. The *alternative forms strategy* involves measuring the N entities at approximately the same time using two forms of a measure that are designed to assess the same underlying construct but that have different items. The *internal consistency method* relies on measuring the N entities on a single occasion with a measure having multiple (k) items.

The reliability of a measure can also be characterized in terms of the *standard error of measurement*, denoted herein as σ_{meas} , which should not be confused with the standard error of the mean (σ_M). Assuming that an attribute of an entity was measured a large number of times with a specific measure of the variable x and that its level remained constant over time, the σ_{meas} would be the standard deviation of scores produced by the measure. Alternatively, the σ_{meas} can be thought of in terms of the standard deviation of scores resulting from measuring a given entity with j measures of x that meet the classical test theory requirements of being mutually parallel. In practice, the estimation of σ_{meas} requires neither (a) the repeated measurement of the entity with a single measure of x nor (b) the measurement of the entity with j parallel measures of x .

Assuming the measurement of x across N measured entities, an estimate of σ_{meas} (denoted hereinafter as $\hat{\sigma}_{\text{meas}}$) can be obtained using estimates of the reliability of the measure (r_{xx}) and the standard deviation of scores produced by the measure (s_x):

$$\hat{\sigma}_{\text{meas}} = s_x \sqrt{1 - r_{xx}}.$$

For example, assume that a researcher measured the quantitative ability of 30 individuals using a 25-item test and that the resulting scores had a mean of 18, standard deviation of 5, and an internal consistency-based estimate of reliability of .80. In this case, $\hat{\sigma}_{\text{meas}}$ would be

$$\begin{aligned} \hat{\sigma}_{\text{meas}} &= s_x \sqrt{1 - r_{xx}} \\ &= 5 \sqrt{1 - .80} \\ &= 2.24. \end{aligned}$$

As should be apparent from the formula for the $\hat{\sigma}_{\text{meas}}$, it can vary between the values of 0 (when $r_{xx} = 1$) and s_x (when $r_{xx} = 0$).

Uses of the Standard Error of Measurement

Assuming that the value of r_{xx} is known, the $\hat{\sigma}_{\text{meas}}$ can be used to establish a confidence interval around an

individual's true (x_t), as opposed to the individual's observed (x_o) score. The estimated value of x_t is given by

$$\hat{x}_t = x_o r_{xx}$$

For example, assume that a person took the above-described quantitative ability test and obtained a score of 15. The individual's estimated true score would be

$$\begin{aligned}\hat{x}_t &= x_o r_{xx} \\ &= 15(.80) \\ &= 12.\end{aligned}$$

Having this estimate, a 95% confidence interval about it can be established by use of

$$\hat{x}_t \pm 1.96\sigma_{\text{meas}}.$$

Thus, the individual's true score would fall in an interval having a lower limit of $7.61 = 12 - 1.96(2.24)$ and an upper limit of $16.39 = 12 + 1.96(2.24)$. In view of the width of the interval, the test would not be very useful for applied purposes, especially if scores on the test were used for important decision-making purposes (e.g., employee hiring). Note, however, that if the same test had a reliability of .90, the confidence interval would extend from 8.90 to 15.10. The general notion here is that the smaller the $\hat{\sigma}_{\text{meas}}$, the narrower the confidence interval.

Importance of Narrow Confidence Intervals

Narrow confidence intervals are important when scores on one variable (e.g., x) are used to predict scores on another (e.g., y). For example, GRE scores are used by most universities to predict success in graduate school. In this and other prediction contexts, the narrower the confidence interval, the greater the degree to which the scores will predict other variables accurately. The reason for this is that nonsystematic variance in scores leads to the attenuation of relations between predictor scores and scores on a predicted variable.

In practice, $\hat{\sigma}_{\text{meas}}$ can be decreased by increasing the reliability of a measure. There are several ways of doing this. In the case of a multi-item measure, two strategies are (a) increasing the number of items in the measure, and (b) including items that have a high degree of correlation with one another.

—Eugene F. Stone-Romero

Further Reading

- Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30, 1–15. Available from www.sportsci.org/resource/stats/precision.html
- Howard, P. (1996, February). *Standard error of measurement (SE_M)*. Technical assistance paper, Florida Department of Education, Division of Public Schools, Bureau of Student Services and Exceptional Education. Retrieved from <http://sss.usf.edu/pdf/standerrmeastap.pdf>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

STANDARD SCORES

A standard score is a score that uses the same metric as other standard scores so that they can be compared to one another easily. Any standard score can be defined as a converted score (based on the raw score) with a set mean and standard deviation. Popular examples are scores on IQ tests and the two types that we will cover here, are called z scores and T scores.

What both of these, as well as other standard scores, have in common is that they share the same mean and standard deviation. It is because of this that standard scores from different distributions can be compared to one another. However, the one restriction is that only the same type of standard scores can be compared. One can compare z scores from one distribution with z scores from another, but one cannot compare z scores from one distribution with T scores from another.

For example, here is a list of 10 raw scores and their corresponding z scores. All the information needed to compute the z scores is contained in Table 1.

Table 1 Raw Scores and Standard Scores

	Raw Score	z Score
	87	0.66
	78	−0.07
	93	1.15
	69	−0.81
	95	1.31
	57	−1.79
	87	0.66
	71	−0.64
	69	−0.81
	68	−0.89
Mean	77.40	
s	12.58	

Note: s = standard deviation.

The z scores are standard scores with a mean of 0 and a standard deviation of 1. The formula for computing a z score is

$$z = \frac{(X - \bar{X})}{s},$$

where

z is the standard z score,

X is the raw score,

$\bar{X}$ is the mean of the distribution, and

s is the standard deviation for the distribution.

For example, where the mean is 77.4 and the standard deviation is 12.58, the z score for a corresponding raw score of 69 is −.81 (almost one standard score below the mean). Here's the formula:

$$z = \frac{(69 - 77.4)}{12.58}.$$

Another popular type of standard score is the *T* score. *T* scores are standard scores with a mean of 50 and a standard deviation of 10. The formula for computing a *T* score is

$$T = 50 + 10z,$$

Table 2 Raw Scores, z Scores, and *T* Scores

	Raw Score	z Score	<i>T</i> Score
	87	0.66	56.61
	78	−0.07	49.27
	93	1.15	61.50
	69	−0.81	41.92
	95	1.31	63.13
	57	−1.79	32.14
	87	0.66	56.61
	71	−0.64	43.56
	69	−0.81	41.92
	68	−0.89	41.11
Mean	77.40		
s	12.58		

Note: s = standard deviation.

where

T is the standard *T* score and

z is the standard z score.

Table 2 shows the listing of *T* scores and z scores for the raw scores shown earlier.

For example, when the mean is 77.4 and the standard deviation is 12.58, the *T* score for a corresponding raw score of 69 is 41.92 (almost one standard score below the mean):

$$T = 50 + 10z \text{ or } 50 - 8.1 \text{ or } 41.9.$$

—Neil J. Salkind

See also Mean; Standard Deviation

STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

The current revision of the *Standards for Educational and Psychological Testing* is the third version of the *Standards* and, like its predecessors, it is the collaborative effort of three prominent national associations

interested in educational and psychological tests: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The first edition of the *Standards* appeared in 1974. It replaced a document published by APA in 1966 and prepared by a committee representing the APA, AERA, and NCME titled *Standards for Educational and Psychological Tests and Manuals*. The 1974 *Standards* was revised in 1985.

To identify revisions needed for the *Standards*, a rigorous and diligent effort to obtain input from the broad measurement community was undertaken. In 1993, the presidents of APA, AERA, and NCME appointed a 15-member Joint Committee with diverse backgrounds in testing and assessment in a variety of settings. The Joint Committee completed the revision during 6 years. Three extended periods of public review and comment provided the Joint Committee with more than 8,000 pages of comments from upwards of 200 organizations and individuals. The Joint Committee considered all this input and developed a draft document. An extensive legal review of this draft was then conducted to explore potential liability issues and to ensure compliance with existing federal law. The revised *Standards* represents a consensus of the Committee and has the endorsement of each of its three sponsoring organizations.

Purpose of the *Standards*

The intent of the third edition of the *Standards* is to promote sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices by providing a frame of reference to assure that all relevant issues are addressed. Like its predecessors, the third edition attempts to reflect professional consensus regarding expectations for the development, validation, and use of educational and psychological tests. The *Standards* is intended to speak broadly to individuals (e.g., students, parents, teachers, administrators, job applicants, employees, clients, patients, supervisors, executives, and evaluators, among others), institutions (e.g., schools, colleges, businesses, industry, clinics, and government agencies), and society as a

whole about tests and testing. The *Standards* can be used to help test publishers decide how to develop, validate, and present tests. The *Standards* can also be used by test users (those who administer tests) to select, use, and evaluate tests. The *Standards* does not attempt to provide psychometric answers to public policy issues that involve testing. Instead, the *Standards* encourages making relevant technical information about tests and testing available so that those involved in policy development may be fully informed.

Organization of the *Standards*

The current revision of the *Standards* contains three parts. Part I addresses test construction, evaluation, and documentation; Part II discusses fairness in testing; and Part III covers testing applications. The standards that apply to the development of tests and those that are of interest to test publishers appear primarily in Part I. The standards identified in Part II and Part III apply primarily, but not exclusively, to test users.

Part I includes the following chapters:

1. Validity
2. Reliability and Errors of Measurement
3. Test Development and Revision
4. Scales, Norms, and Score Comparability
5. Test Administration, Scoring, and Reporting
6. Supporting Documentation for Tests

Part II includes the following chapters:

7. Fairness in Testing and Test Use
8. The Rights and Responsibilities of Test Takers
9. Testing Individuals of Diverse Linguistic Backgrounds
10. Testing Individuals with Disabilities

Part III includes the following chapters:

11. The Responsibilities of Test Users
12. Psychological Testing and Assessment

13. Educational Testing and Assessment
14. Testing in Employment and Credentialing
15. Testing in Program Evaluation and Public Policy

Each chapter begins with contextual background intended to facilitate interpretation and application of the standards in that chapter. An index and a glossary that defines terms as they are used in the *Standards* also are provided.

Major Differences Between the Second and Third Editions of the *Standards*

The overall number of standards has increased from the 1985 edition for three reasons. First, new types of tests and uses for tests evolved after the 1985 revision. Several of the new standards apply only to these new developments, unlike the broad applicability that is characteristic of many of the continuing standards. Second, some standards are repeated, with context-relevant wording changes, in multiple chapters to accommodate users who refer only to those chapters that have direct relevance to their particular setting or purpose for testing. The wording changes enable standards to align with the content of the chapter. Third, standards addressing issues such as conflict of interest and equitable treatment of test takers have been added. According to the *Standards*, “The increase in the number of standards does not per se signal an increase in the obligations placed on test developers and test users.”

The 1985 *Standards* categorized each standard as either primary (to be met by all tests before they are used), secondary (desirable but not feasible in all situations), or conditional (importance varies with application). This distinction was eliminated for the third edition of the *Standards* because it was recognized that the various standards are not absolutes. The applicability of the standards can vary in relevance and importance based on the intended use(s) of a test and the role of the person(s) participating in the testing process (e.g., test taker, test administrator, test developer, test marketer, and those who make decisions

based on test results). The third edition also clarifies that some standards are broad and encompassing in their applicability and others are narrower in scope. Therefore, the standards should not be applied in a literal, rigid, or cookbook fashion. Instead, whether a test or use is judged acceptable may depend on several interrelated and interacting factors. These can include (a) the extent to which the test developer has met relevant standards, (b) the degree to which the test user has met relevant standards, (c) the availability of alternative, equally promising measures, (d) the extent of empirical research that supports the intended use of the measure, and (e) professional judgment. The *Standards* advises that before a test is operationally used for a particular purpose, “each standard should be carefully considered to determine its applicability to the testing context under consideration.”

Each chapter in the third edition contains more introductory material than did the second edition. The purpose of this additional information is to provide background for the standards specific to the chapter so that users can more easily interpret and apply the standards. The language, although prescriptive at times, “should not be interpreted as imposing additional standards.”

The third edition defines and clarifies several important terms. For example, the term *test* is defined as an “evaluative device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process.” Thus the term *test* is broad, including both instruments that are evaluated for quality or correctness and instruments that are measures of attitudes, interests, traits, and dispositions, often referred to as scales or inventories rather than tests. Assessment is considered to be a broader concept than testing, but testing is part of assessment. *Assessment* refers to a process that integrates test information with background and contextual information, whereas *testing* refers to the results obtained from a specific instrument or instruments.

The new *Standards* broadens the meaning of the term *construct*. In previous editions, *construct* meant unobservable characteristics that must be inferred from multiple, related observations, a definition that

proved confusing and controversial. The third edition broadens the term to mean the “concept or characteristic that a test is designed to measure.” This change requires test professionals to specify the interpretation of the construct that will be made on the basis of a score or pattern of scores. This change also reflects a shift in the third edition of the *Standards* from discussing types of validity to discussing various *lines of validity evidence* that serve to enhance interpretation of a score relative to the construct the test is designed to measure.

Lines of Validity Evidence

Formerly, the *Standards* described three types of validity that a test may demonstrate: content, criterion-related (predictive and concurrent), and construct. The new *Standards* considers validity to be the extent to which multiple lines of evidence and theory “support the interpretations of test scores entailed by the proposed uses of tests.” According to the new *Standards*, then, validity is a function of the extent to which theory and empirical evidence support the assumption that a test score reflects the construct the test purports to measure. Five sources of such evidence of validity are identified in the *Standards*: test content, response processes, internal structure, relations to other variables, and consequences of testing.

Of these five sources of validity, three reflect historical conceptions of validity that have appeared in previous editions of the *Standards*. These are test content (equivalent to content validity), internal structure (equivalent to construct validity), and relations to other variables (equivalent to criterion-related validity). The two new sources of validity evidence are response process and consequences of testing. Evidence of response process validity indicates whether the test taker used the processes intended by the test developer, rather than an unintended process, to respond to a problem. For example, in responding to a mathematics problem, did the test taker apply the intended cognitive-mathematical process or an alternative process, such as guessing? Validity evidence related to the consequences of testing emerges when

the outcome and interpretation of the test fulfill the claims made for the test. For example, if a measure purports to predict those who will benefit from a particular psychological treatment, how well does it actually do so?

A number of advances and developments have occurred in testing since the *Standards* was released in 1999. Thus, this edition of the *Standards* should be considered a work in progress. AERA, APA, and NCME have already begun the revision process for the fourth edition. Although an exact publication date cannot be determined, it is expected to come early in the next decade.

—Thomas Kubiszyn

Further Reading

- American Educational Research Association. (1985). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: Author.
- American Educational Research Association. (1999). *Standards for educational and psychological testing* (3rd ed., pp. 3–9). Washington, DC: Author.
- Braden, J. P., & Niebling, B. C. (2005). Using the joint test standards to evaluate the validity evidence for intelligence tests. In D. F. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed.). New York: Guilford.

Standards for Educational and Psychological Testing: <http://www.apa.org/science/standards.html>

STANFORD ACHIEVEMENT TEST

The Stanford Achievement Test Series (publisher: Harcourt Assessment) is a set of 13 norm-referenced grade-level assessments of academic achievement in mathematics, reading, language, spelling, science, social science, and listening. The Stanford Early School Achievement Test levels are designed to assess initial school learning across kindergarten and the first half of first grade. The Stanford Achievement Test (Primary 1–3, Intermediate 1–3, and Advanced 1–2) spans the second half of first grade through ninth grade. Finally, the three-level Stanford Test of

Academic Skills assesses performance on curricula in Grades 9 through 12.

This typically group-administered test is available in a number of formats, including general and secure forms, full-length and abbreviated batteries of multiple choice items, and large-print and Braille editions. Also, special screening procedures for assigning deaf and hard of hearing students to test levels have been developed by the Gallaudet Research Institute.

The test score metrics available to assist with the interpretation of student test performance include individual percentile rank, normal curve equivalent, stanine, grade equivalent, content cluster and process cluster performance category, and scaled score. The scaled score, derived from the vertical equating of test levels, is particularly useful for monitoring individual student achievement test results from year to year.

The design and materials for the latest (10th) edition of the Stanford were developed to conform to the 1999 AERA, APA, and NCME *Standards for Educational and Psychological Testing*. The design strategy emphasized universal design principles and visual characteristics that are appealing and resemble everyday instructional materials. Such features include full color “lifelike” illustrations, visually distinct framing of each question, placement of response bubbles on answer sheets to correspond with page and subtest location in the test booklet, and the absence of time constraints (although suggested testing times are given). Item development was responsive to widely promulgated national curriculum standards. To satisfy individual state academic standards, a core subset of items from the Stanford is augmented with specifically developed items on a state-by-state contract basis.

The Stanford was first published in 1923 by the World Book Company and is now published by Harcourt Assessment, the testing division of Reed Elsevier, which acquired Harcourt in 2001. The Stanford was developed for a longitudinal study of gifted children under the direction of Lewis M. Terman (1877–1956) and was first offered at only two elementary grade levels. The most recent update, the 10th edition, was normed in the spring and fall of 2002 and published in 2003. Regarded as the first standardized academic achievement test, the Stanford

has remained a highly respected assessment instrument with excellent psychometric characteristics.

—Ross E. Mitchell and Michael A. Karchmer

Further Reading

Carney, R. N., & Morse, D. T. (2005). [Reviews of the Stanford Achievement Test, 10th Edition]. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 969–975). Lincoln, NE: Buros Institute of Mental Measurements.

Harcourt Assessment: http://www.harcourt.com/bu_info/harcourt_assessment.html

STANFORD-BINET INTELLIGENCE SCALES

The Stanford-Binet Intelligence Scales is one of the most popularly used measures of intelligence. Around 1905, Parisian Theophilus Simon, in his work with developmental psychologist Jean Piaget, speculated about the successes and failures of elementary school children. As Piaget was becoming increasingly intrigued by the nature of some normal children’s failures to read, Simon was working with Alfred Binet to develop a test (for the school children of Paris) to predict and distinguish between those who would do well in school and those who would not.

The result of Simon and Binet’s efforts was the first intelligence scale (1905) used to identify children who were mentally retarded (or those who were thought not to be able to succeed in school). The items were arranged by difficulty at first and later on by age level, so by 1916, the test results yielded a *mental age* (MA) for test takers, and a ratio of chronological age (CA) and mental age provided a measure of whether a child was behind, even with, or ahead of his or her chronological age. For example, someone who is chronologically 120 months (10 years) old may have a mental age of 126 months and therefore be advanced in mental age, compared with chronological age.

The Simon-Binet scale became the Stanford-Binet scale in the early 20th century, when Lewis Terman

(then at Stanford University) authored the authoritative book on the administration and scoring of the test. Also, the IQ score, a ratio of mental age to chronological age, was first used. For the 120-month-old with a mental age of 126 months,

$$IQ = \frac{MA}{CA} \times 100 = \frac{126}{120} \times 100 = 105.$$

If one's mental and chronological ages are equal, then one's IQ score is 100, just about what we expect if someone is average.

It was convenient to have one number, a simple ratio of MA to CA, to express one's theoretical intelligence quotient. But this approach was never really embraced because the upper limit depends on the upper age limit for the test. For example, if the items go up only to an age level of 21 years, anyone with a CA greater than 21 years has to have an IQ less than 100. This is in part why we no longer think of intelligence as IQ but use more descriptive and informative terms for describing someone's level of intelligence.

The fourth edition of the Stanford-Binet is now the standard, and it is the result of many revisions over the past 100 years. The test, based on the *g* theory of intelligence, assesses three different general areas or types of intelligence: crystallized intelligence, fluid intelligence, and short-term memory. In theory, these types are independent of one another.

Crystallized intelligence reflects knowledge that is acquired or learned, such as the number of elements on the periodical table or the author of *Moby Dick*. Fluid intelligence is that general ability that Charles Spearman talked about, including such activities as problem solving, remembering, and learning. A test item that evaluates fluid intelligence might ask the test taker to copy a pattern created by the test administrator.

Today's Stanford-Binet assesses five different factors, fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory—outgrowths of the fluid, crystallized, and memory typology. Each of these areas is assessed verbally and nonverbally, resulting in 10 subtests. In the end, each individual test taker has a score for verbal and nonverbal performance and a full-scale intelligence score.

The Stanford-Binet assesses intelligence by beginning with those items that are the easiest and then finding out which items become so difficult that there is no need to test further. This range of difficulty helps define the starting and ending points in the scoring of performances.

The basal age is the lowest point on the test at which the test taker can pass two consecutive items that are of equal difficulty. The examiner can feel confident that the test taker is on firm ground at this point and could pass all the items that are less difficult. Ceiling age is the point at which at least three out of four items are missed in succession and is the place to stop testing.

The number correct on the test is then used to compute a raw score, which is converted to a standardized score and then compared with other scores from test takers in the same age group. Because the Stanford-Binet is a standardized test and has been normed extensively, such comparisons are easy to make and very useful.

—Neil J. Salkind

Further Reading

Alfred Binet and his contributions: <http://www.indiana.edu/~intell/binet.shtml>

Stanford-Binet Intelligence Scales: <http://www.riverpub.com/products/sb5/index.html>

STANINE

The term *stanine* comes from the field of measurement, for a person's performance on a test can be (and often is) converted into a stanine score. An examinee's stanine score shows the examinee's position relative to other test takers' positions. Because of this focus on relative performance, stanines are similar to percentile ranks, *T* scores, or normal curve equivalent (NCE) scores. Though like these other kinds of standard scores in purpose, stanines have many characteristics that make them unique and quite different from the others.

A person's stanine score will be a single-digit number, and it could be as high as 9, as low as 1, or equal

Table 1 Stanines and Their Respective Percentage Groupings

Stanine Score	% Reference Group in Each Stanine Category
9	4
8	7
7	12
6	17
5	20
4	17
3	12
2	7
1	4

to any whole number in between. In the group of test takers that serves as the *reference group*, fewer individuals receive the higher and lower stanine scores, and most people earn stanine scores at or near the middle of the 9-point scale. Table 1 shows what percentage of individuals in the reference group receives each stanine score.

Although stanine scores are fairly easy to understand from the information in Table 1, they can be made even easier to understand by attaching verbal descriptors to the individual stanine scores or to groups of adjacent stanine scores. Some people recommend dividing the stanine scale into three parts and using the terms *above average*, *average*, and *below average* to describe those whose test scores position them in any of the top three, middle three, or bottom three stanine categories, respectively. Others

recommend using these descriptors: *outstanding* for stanine 9, *above average* for stanines 7 and 8, *average* for stanines 4, 5, and 6, *below average* for stanines 2 and 3, and *poor* for stanine 1. A few people recommend using a different label for each stanine category, as indicated in Table 2.

The Meaning and Origin of the Word *Stanine*

The word *stanine* is an abbreviation of the phrase *standard of nine*. With stanines, there is a standard of nine because the reference group is always divided into nine categories, or levels, with the specific percentages set equal to 4, 7, 12, 17, 20, 17, 12, 7, and 4. It doesn't matter whether the reference group is bright or dull, hard-working or lazy; the reference group's top 4% will end up getting a stanine score of 9, the next best 7% will receive a stanine score of 8, and so forth. Thus, the percentages in the various stanine categories are constant regardless of the characteristics of the reference group.

The term *stanine* was first used by the U.S. Army Air Force during World War II. Recruits were given a battery of placement tests, and each examinee's performance (on each portion of the test battery) was converted into a stanine score. At the time, having test scores converted into single-digit numbers was especially helpful, for each examinee's data were entered into a computer. Back then, however, data were typed on Hollerith punch cards, and it was far more efficient during data entry to use one column per score (without the need for any decimal points or negative signs).

Table 2 Stanines and Their Respective Labeling

Stanine Score	Label
9	high
8	well above average
7	above average
6	somewhat above average
5	about average
4	somewhat below average
3	below average
2	well below average
1	low

Advantages of Stanines Over Raw Scores

After a test is administered and the examinees' work is scored, each test taker usually earns a raw score based on the number of questions answered correctly. Raw scores are useless by themselves. For example, if Johnny gets a 43 on a math test, there is no way to know from this raw score how well (or poorly) Johnny performed. If Johnny's raw score is converted into a percentage-correct score or compared to a cut-score

for passing the test, we get a better feel for how Johnny did. However, simply knowing Johnny's percentage-correct score or whether he did or did not pass the test provides no information whatsoever as to how Johnny's performance compares with that of others who took the same test.

When an examinee's raw score is converted into a stanine score, we learn how that particular test taker compares with other test takers. There are different ways to convert a raw score into some other score that indicates, in a relative sense, how well or poorly a person has done when taking a test. Later in this entry, stanines will be compared with percentile ranks, *T* scores, and NCE scores.

Stanines, Norm-Referenced Tests, and Criterion-Referenced Tests

As indicated earlier, stanines provide information as to how a test taker performs relative to others in a reference group. Because of this inherent feature of stanine scores, they are used exclusively with norm-referenced tests. They are *not* to be used with criterion-referenced tests. With this latter kind of test, comparisons against others are of little or no interest to those who deal with the scores examinees earn.

Converting Raw Scores Into Stanine Scores

There are two ways to convert an examinee's raw score into a stanine score. One of these methods can

be called the *ranking* procedure. The other method involves reference to a normal distribution, so this second method might be called the *normal distribution* procedure.

The ranking method for determining stanine scores is quite simple and straightforward. After the examinees' raw scores are determined, those scores are ranked from high to low. Then, the top 4% are given a stanine score of 9, the next highest 7% are given a stanine score of 8, and so on. Note that this method of creating stanine scores makes no assumption as to the distributional shape of the raw scores.

The second procedure for determining stanine scores is used with tests that have been normed on a large group of individuals who have already taken the test and whose scores have been normalized. A normal distribution is divided into nine sections, with the upper and lower boundaries of each stanine section set equal to the *z* scores shown in Table 3.

Except for stanines 1 and 9, each stanine section has a width equal to one half of a standard deviation. Because a normal distribution is considered to have no upper or lower scores, stanine 9 has a boundary of $z = +1.75$ on the left but no boundary on the right. Similarly, stanine 1 has a boundary on the right of $z = -1.75$ but no boundary on the left.

Because of the known properties of a normal distribution, it is possible to define the boundary lines between adjacent stanine categories in terms of percentiles rather than *z* scores. Rounded to whole numbers, these percentile boundaries are shown in Table 4.

Table 3 Stanines *z* Scores

Stanines Located on the Left Side of the Normal Distribution			The Middle Stanine (i.e., Stanine 5)		Stanines Located on the Right Side of the Normal Distribution		
Stanine section	Lower boundary	Upper boundary	Lower boundary	Upper boundary	Stanine section	Lower boundary	Upper boundary
4	$z = -.75$	$z = -.25$	$z = -.25$	$z = +.25$	6	$z = +.25$	$z = +.75$
3	$z = -1.25$	$z = -.75$			7	$z = +.75$	$z = +1.25$
2	$z = -1.75$	$z = -1.25$			8	$z = +1.25$	$z = +1.75$
1	No lower boundary	$z = -1.75$			9	$z = +1.75$	No upper boundary

Table 4 Percentile Boundaries

<i>Adjacent stanines</i>	1 and 2	2 and 3	3 and 4	4 and 5	5 and 6	6 and 7	7 and 8	8 and 9
<i>Percentile of boundary</i>	4	11	23	40	60	77	89	96

Each of these percentiles, of course, can be determined by adding together the stanine percentages for those stanines located to the left of the boundary line being considered. For example, the percentile boundary separating stanines 5 and 6 can be determined by adding 4, 7, 12, 17, and 20.

How Stanines Are Used

Stanines are used in school settings to accomplish three different goals. In two of these uses, the focus is on the individual test taker. When stanines are used to accomplish the third goal, the focus is a group of test takers rather than a single person.

If a test yields just one score or a “total” score based on several parts of the test, that score often is converted into a stanine to show how the test taker performed relative to others in the reference group. For example, if the examinee’s stanine score is 8, that examinee would be thought of as having a relatively high position within the reference group. On the other hand, a stanine of 1 would indicate that an examinee was positioned quite low compared to others who took the same test. Stanine scores such as these are often used to place students into homogeneous learning groups.

If a test yields multiple scores, each for a different skill or competency being measured, a test taker’s resulting scores can each be converted into a stanine. Then, the individual’s various stanine scores can be examined so as to identify areas of strength and areas of weakness. For example, if a pupil earns a stanine score of 8 in reading, a score of 7 in writing, a score of 9 in spelling, and a score of 4 in arithmetic, such scores can be used to tailor instruction to the places where the pupil is “ahead” or “behind.”

Stanine scores are also used within statements of instructional goals. For example, a magnet elementary school recently articulated this goal for its third-grade

teachers: “In the area of math, the percentage of students scoring in the high stanines will be 65% or higher and no more than 5% scoring in the low stanines as measured on the [Iowa Test of Basic Skills].” In this third use of stanine scores, the focus is on a group of students rather than on individual pupils.

Stanines Versus Other Kinds of Standard Scores

Stanine scores are easier to understand than other kinds of standard scores, such as z scores, T scores, or NCE scores. This is because (a) there are only nine stanine scores, (b) no decimals or negative values are involved, and (c) people have little difficulty “catching on” to the fact that most examinees end up in stanines 4, 5, or 6; that fewer test takers end up in stanines 3 and 4 (or stanines 7 and 8); and that only a very small percentage of examinees end up in stanines 1 or 9. Thus, it is not surprising that stanine scores are used in parent-teacher conferences and in reports that summarize students’ test performance.

Although stanines have certain distinct advantages over other standard scores, they have one main limitation. Simply stated, the difference between two people can be completely hidden when their test performance is reported via stanines, or conversely, two people who are very similar may end up “looking dissimilar” because of their stanine scores. To illustrate this disadvantage of stanines, suppose four individuals take a test and perform such that their percentile ranks are 41, 59, 76, and 78. The first two of these test takers would both have a stanine score of 5 despite the difference in how they performed on the test. The third and fourth test takers would end up in different stanines despite the similarity in how they performed. These two undesirable situations would not occur if the performance of these four test takers is converted into z scores, T scores, or NCE scores.

Warnings About Stanine Scores

For those who use stanine scores, it is important to keep four things in mind. First, we should always know who formed the reference group when we attempt to interpret an examinee's stanine score. If the reference group was quite bright or highly skilled, an examinee thought to be average (or even above average) in ability could end up with a low stanine score. Second, stanine scores should not be averaged. Third, a comparison of two stanine scores can be misleading because of the way test scores are converted into stanines. This third warning applies to the comparison of two different examinees or to the comparison of one examinee's standing at two different times.

Finally, those who use stanine scores should always be concerned about issues of reliability and validity. Without much difficulty, one could create a test, administer that test to a large reference group, and then create a conversion table that allows test scores to be easily transformed into stanine scores. Yet we must ask the critical question whether a new test taker's stanine score can be trusted to mean what we are tempted to think it does. As with other ways of reporting test performance (such as percentile ranks or *T* scores), stanine scores should be interpreted only if they are tied to a measuring instrument that has known (and respectable) psychometric properties.

—Schuyler W. Huck

See also Percentile and Percentile Rank; Standard Scores

Further Reading

- Allen, M. J., & Yen, W. M. (1979). *Introduction to test theory*. Monterey, CA: Brooks-Cole.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Upper Saddle River, NJ: Prentice Hall.
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.

Stanine article: <http://en.wikipedia.org/wiki/Stanine>

STATIS

STATIS is a generalization of principal component analysis (PCA), and its goal is to analyze several sets of variables collected on the same set of observations. It is attributed to Escouffier and L'Hermier des Plantes. A related approach is known as *procrustes matching by congruence coefficients* in the English-speaking community. The goal of STATIS is (a) to compare and analyze the relationship between the different data sets, (b) to combine them into a common structure called a *compromise*, which is then analyzed via PCA to reveal the common structure between the observations, and finally (c) to project each of the original data sets onto the compromise to analyze communalities and discrepancies. STATIS is used in very different domains, such as sensory evaluation, molecular imaging, brain imaging, ecology, and chemometrics.

The number or nature of the variables used to describe the observations can vary from one data set to the other, but the observations should be the same in all the data sets.

For example, the data sets can be measurements taken on the same observations (individuals or objects) at different occasions. In this case, the first data set corresponds to the data collected at time $t = 1$, the second to the data collected at time $t = 2$, and so on. The goal of the analysis, then, is to evaluate whether the position of the observations is stable over time.

As another example, the data sets can be measurements taken on the same observations by different participants or groups of participants. In this case, the first data set corresponds to the first participant, the second to the second participant, and so on. The goal of the analysis, then, is to evaluate whether there is an agreement between the participants or groups of participants.

The general idea behind STATIS is to analyze the structure of the individual data sets (i.e., the relation between the individual data sets) and to derive from this structure an optimal set of weights for computing a common representation of the observations, called the *compromise*. The weights used to compute the compromise are chosen to make it as representative of

all the data sets as possible. The PCA of the compromise gives, then, the position of the observations in the compromise space. The position of the observations for each data set can be represented in the compromise space as supplementary points. Finally, as a by-product of the weight computation, the data sets can be represented as points in a multidimensional space.

An Example

A typical example of using STATIS is the description of a set of products by a group of experts. This type of data can be analyzed using a standard PCA after the data have been averaged across experts. However, this approach neglects the inter-expert differences. STATIS has the advantages of providing a compromise space for the products as well as evaluating the differences among experts. We illustrate the method with an example from wine tasting.

Red wines often spend several months in oak barrels before being bottled because oak adds interesting components to the wine. However, only certain species of oaks seem to work well. Suppose we wanted to evaluate the effect of the oak species on barrel-aged red burgundy wines. First, we selected six wines coming from the same harvest of pinot noir and aged in six different barrels made with one of two different types of oak. Wines 1, 5, and 6 were aged with the first type of oak, whereas wines 2, 3, and 4 were aged with the second. Next, we asked each of three wine experts to choose from two to five variables to describe the six wines. For each wine, each expert was asked to rate the intensity of the variables on a 9-point scale. The results are presented in Table 1. The goal of

the analysis is twofold. First, we want to obtain a typology of the wines, and second, we want to know whether there is agreement among the experts.

Notations

The raw data consist of T data sets. For convenience, we will refer to each data set as a study. Each study is an $I \times J_{[t]}$ rectangular data matrix denoted $\mathbf{Y}_{[t]}$, where I is the number of observations and $J_{[t]}$ the number of variables collected on the observations for the t th study. Each data matrix is, in general, pre-processed (e.g., centered, normalized), and the pre-processed data matrices actually used in the analysis are denoted $\mathbf{X}_{[t]}$.

For our example, the data consist of $T = 3$ studies. The data were centered by column (i.e., the mean of each column is zero), and the starting point of the analysis consists of three matrices $\mathbf{X}_{[t]}$:

$$\mathbf{X}_{[1]} = \begin{bmatrix} -3.00 & 2.67 & 3.67 \\ 1.00 & -0.33 & -1.33 \\ 2.00 & -2.33 & -2.33 \\ 3.00 & -2.33 & -1.33 \\ -2.00 & 1.67 & 0.67 \\ -1.00 & 0.67 & 0.67 \end{bmatrix}, \quad (1)$$

$$\mathbf{X}_{[2]} = \begin{bmatrix} -2.00 & 1.17 & 3.17 & 2.50 \\ 0.00 & 0.17 & 0.17 & -1.50 \\ 1.00 & -1.83 & -2.83 & -2.50 \\ 3.00 & -1.83 & -2.83 & -1.50 \\ -1.00 & 1.17 & 2.17 & 1.50 \\ -1.00 & 1.17 & 0.17 & 1.50 \end{bmatrix},$$

Table 1 Raw Data for the Barrel-Aged Red Burgundy Wines Example

Wines	Oak-type	Expert 1			Expert 2				Expert 3		
		fruity	woody	coffee	red fruit	roasted	vanillin	woody	fruity	butter	woody
wine1	1	1	6	7	2	5	7	6	3	6	7
wine2	2	5	3	2	4	4	4	2	4	4	3
wine3	2	6	1	1	5	2	1	1	7	1	1
wine4	2	7	1	2	7	2	1	2	2	2	2
wine5	1	2	5	4	3	5	6	5	2	6	6
wine6	1	3	4	4	3	5	4	5	1	7	5

$$\text{and } \mathbf{X}_{[3]} = \begin{bmatrix} -0.17 & 1.67 & 3.00 \\ 0.83 & -0.33 & -1.00 \\ 3.83 & -3.33 & -3.00 \\ -1.17 & -2.33 & -2.00 \\ -1.17 & 1.67 & 2.00 \\ -2.17 & 2.67 & 1.00 \end{bmatrix}.$$

Each of the $\mathbf{X}_{[i]}$ matrices is then transformed into an $I \times I$ scalar product matrix denoted $\mathbf{S}_{[i]}$ and computed as

$$\mathbf{S}_{[i]} = \mathbf{X}_{[i]} \mathbf{X}_{[i]}^T. \quad (2)$$

For example, the 6×6 between-wine scalar product matrix for the first wine expert is denoted $\mathbf{S}_{[1]}$. It is obtained as

$$\mathbf{S}_{[1]} = \mathbf{X}_{[1]} \mathbf{X}_{[1]}^T = \begin{bmatrix} 29.56 & -8.78 & -20.78 & -20.11 & 12.89 & 7.22 \\ -8.78 & 2.89 & 5.89 & 5.56 & -3.44 & -2.11 \\ -20.78 & 5.89 & 14.89 & 14.56 & -9.44 & -5.11 \\ -20.11 & 5.56 & 14.56 & 16.22 & -10.78 & -5.44 \\ 12.89 & -3.44 & -9.44 & -10.78 & 7.22 & 3.56 \\ 7.22 & -2.11 & -5.11 & -5.44 & 3.56 & 1.89 \end{bmatrix}. \quad (3)$$

The scalar product matrices for the second and third wine experts are equal to

$$\mathbf{S}_{[2]} = \begin{bmatrix} 21.64 & -3.03 & -19.36 & -20.86 & 13.97 & 7.24 \\ -3.03 & 2.31 & 2.97 & 1.47 & -1.69 & -2.03 \\ -19.36 & 2.97 & 18.64 & 18.14 & -13.03 & -7.36 \\ -20.86 & 1.47 & 18.14 & 22.64 & -13.53 & -7.86 \\ 13.97 & -1.69 & -13.03 & -13.53 & 9.31 & 4.97 \\ 7.64 & -2.03 & -7.36 & -7.86 & 4.97 & 4.64 \end{bmatrix}. \quad (4)$$

and

$$\mathbf{S}_{[3]} = \begin{bmatrix} 11.81 & -3.69 & -15.19 & -9.69 & 8.97 & 7.81 \\ -3.69 & 1.81 & 7.31 & 1.81 & -3.53 & -3.69 \\ -15.19 & 7.31 & 34.81 & 9.31 & -16.03 & -20.19 \\ -9.69 & 1.81 & 9.31 & 10.81 & -6.53 & -5.69 \\ 8.97 & -3.53 & -16.03 & -6.53 & 8.14 & 8.97 \\ 7.81 & -3.69 & -20.19 & -5.69 & 8.97 & 12.81 \end{bmatrix}. \quad (5)$$

Computing the Compromise Matrix

The *compromise matrix* is a scalar product matrix that gives the best compromise (hence its name) of the scalar product matrices representing each study. It is obtained as a weighted average of the study scalar product matrices. The weights are chosen so that studies agreeing the most with other studies will have the larger weights. To find these weights, we

need to analyze the relationships between the studies.

Comparing the Studies

To analyze the similarity structure of the studies, we start by creating a *between study cosine matrix* denoted $\mathbf{C}$. This is a $T \times T$ matrix whose generic term

$c_{t,t'}$ gives the cosine between studies. This cosine, also known as the R_V coefficient, is defined as

$$R_V = [c_{t,t'}] = \frac{\text{trace} \{ \mathbf{S}_{[t]}^T \mathbf{S}_{[t']} \}}{\sqrt{\text{trace} \{ \mathbf{S}_{[t]}^T \mathbf{S}_{[t]} \} \times \text{trace} \{ \mathbf{S}_{[t']}^T \mathbf{S}_{[t']} \}}}. \quad (6)$$

Using this formula we get the following matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} 1.00 & .95 & .77 \\ .95 & 1.00 & .82 \\ .77 & .82 & 1.00 \end{bmatrix}. \quad (7)$$

PCA of the Cosine Matrix

The eigendecomposition of the cosine matrix reveals the structure between the studies. This amounts to performing a *noncentered* PCA of $\mathbf{C}$. Formally, this matrix has the following eigendecomposition:

$$\mathbf{C} = \mathbf{P}\mathbf{\Theta}\mathbf{P}^T \text{ with } \mathbf{P}^T\mathbf{P} = \mathbf{I}, \quad (8)$$

where $\mathbf{P}$ is the matrix of eigenvectors of $\mathbf{C}$ and $\mathbf{\Theta}$ the diagonal matrix of eigenvalues. An element of a given eigenvector represents the projection of one study on this eigenvector. Thus the studies can be represented as points in the eigenspace and their similarities visually analyzed. In this case, the projections are computed as

$$\mathbf{G} = \mathbf{P}\mathbf{\Theta}^{\frac{1}{2}}. \quad (9)$$

For our example, we find that

$$\mathbf{P} = \begin{bmatrix} 0.58 & -0.49 & 0.65 \\ 0.59 & -0.29 & -0.75 \\ 0.55 & 0.82 & 0.12 \end{bmatrix} \quad (10)$$

and
$$\mathbf{\Theta} = \begin{bmatrix} 2.70 & 0.00 & 0.00 \\ 0.00 & 0.26 & 0.00 \\ 0.00 & 0.00 & 0.04 \end{bmatrix}$$

and
$$\mathbf{G} = \begin{bmatrix} 0.96 & 0.25 & 0.14 \\ 0.98 & 0.14 & -0.16 \\ 0.91 & -0.42 & 0.03 \end{bmatrix}. \quad (11)$$

As an illustration, Figure 1 displays the projections of the experts onto the first and second components. It shows that the three studies are positively correlated with the first component (this is because all the elements of the cosine matrix are positive).

Computing the Compromise

The weights used for computing the compromise are obtained from the PCA of the cosine matrix. Because this matrix is not centered, the first eigenvector of $\mathbf{C}$ represents what is common to the different studies. Thus studies with larger values on the first eigenvector are more similar to the other studies and therefore will have a larger weight. Practically, the weights are obtained by rescaling the elements of the first eigenvector of $\mathbf{C}$ so that their sum is equal to 1. We call the weight vector α . For our example, we find that

$$\alpha = [.337 \ .344 \ .319]^T. \quad (12)$$

So with α_t denoting the weight for the t th study, the compromise matrix, denoted $\mathbf{S}_{[+]}$, is computed as

$$\mathbf{S}_{[+]} = \sum_t^T \alpha_t \mathbf{S}_{[t]}. \quad (13)$$

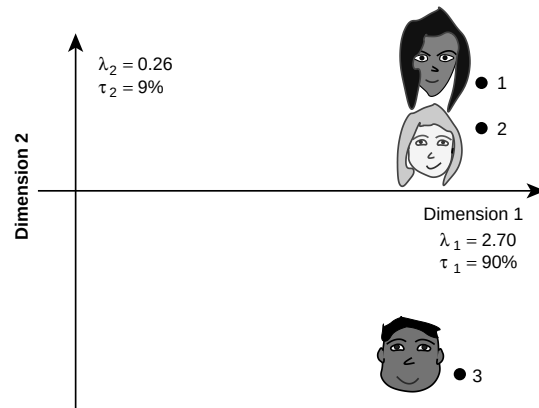


Figure 1 The Expert Space

In our example, the compromise is obtained as

$$\mathbf{S}_{[+]} = \begin{bmatrix} 22.15 & -5.42 & -19.37 & -17.83 & 12.57 & 7.90 \\ -5.42 & 2.45 & 5.59 & 3.09 & -3.01 & -2.71 \\ -19.37 & 5.59 & 23.61 & 14.76 & -13.38 & -11.21 \\ -17.83 & 3.09 & 14.76 & 17.47 & -10.84 & -6.65 \\ 12.57 & -3.01 & -13.38 & -10.84 & 8.61 & 6.05 \\ 7.90 & -2.71 & -11.21 & -6.65 & 6.05 & 6.62 \end{bmatrix}. \quad (14)$$

How Representative Is the Compromise?

The compromise is the best aggregate of the original scalar product matrices. But how good is this “best”? An index of the quality of the compromise is given by the ratio of the first eigenvalue of $\mathbf{C}$ to the sum of the eigenvalues of $\mathbf{C}$:

$$\text{Quality of compromise} = \frac{\vartheta_1}{\sum_{\ell} \vartheta_{\ell}} = \frac{\vartheta_1}{\text{trace}\{\mathbf{\Theta}\}}. \quad (15)$$

For our example, the quality of the compromise is evaluated as $\frac{2.74}{3} \approx .91$. So we can say that the compromise “explains” 91% of the inertia of the original set of data tables.

Analyzing the Compromise

The compromise matrix is a scalar product matrix, and therefore its eigendecomposition amounts to a PCA. From this analysis, we can explore the structure of the set of observations. The eigendecomposition of the compromise gives

$$\mathbf{S}_{[+]} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \quad (16)$$

with

$$\mathbf{Q} = \begin{bmatrix} 0.54 & -0.34 & -0.57 & -0.03 & 0.31 \\ -0.14 & -0.15 & 0.49 & 0.54 & 0.51 \\ -0.55 & -0.57 & 0.00 & -0.40 & -0.23 \\ -0.45 & 0.59 & -0.49 & 0.19 & 0.01 \\ 0.34 & 0.03 & 0.17 & 0.34 & -0.75 \\ 0.25 & 0.43 & 0.39 & -0.63 & 0.16 \end{bmatrix} \quad (17)$$

and

$$\text{diag}\{\mathbf{\Lambda}\} = \begin{bmatrix} 69.70 \\ 7.35 \\ 2.52 \\ 1.03 \\ 0.32 \end{bmatrix}. \quad (18)$$

From Equations 17 and 18, we can compute the compromise factor scores for the wines as

$$\mathbf{F} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}} \quad (19)$$

$$= \begin{bmatrix} 4.52 & -0.92 & -0.91 & -0.03 & 0.17(36) \\ -1.13 & -0.42 & 0.78 & 0.54 & 0.29(37) \\ -4.59 & -1.53 & 0.01 & -0.40 & -0.13(38) \\ -3.77 & 1.61 & -0.78 & 0.19 & 0.01(39) \\ 2.87 & 0.08 & 0.27 & 0.34 & -0.43(40) \\ 2.10 & 1.18 & 0.63 & -0.64 & 0.09(41) \end{bmatrix}. \quad (20)$$

In the $\mathbf{F}$ matrix, each row represents an observation (i.e., a wine), and each column is a component. Figure 2 displays the wines in the space of the first two principal components. The first component has an eigenvalue equal to $\lambda_1 = 69.70$, which corresponds to 85% of the inertia because

$$\frac{69.70}{69.70 + 7.35 + 2.52 + 1.03 + 0.32} = \frac{69.70}{80.91} \approx .85.$$

The second component, with an eigenvalue of 7.35, explains almost 10% of the inertia. The first

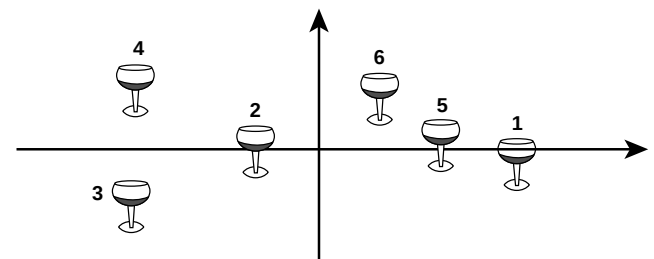


Figure 2 Analysis of the Compromise: Plot of the Wines on the First Two Principal Components

component is easily interpreted as the opposition of the wines aged with the first type of oak (wines 1, 5, and 6) to the wines aged with the second type of oak (wines 2, 3, and 4).

Projecting the Studies Into the Compromise Space

The analysis of the compromise reveals the structure of the wine space common to the experts. In addition, we want to see how each expert “interprets” this space. This is achieved by projecting the scalar product matrix of each expert onto the compromise. This operation is performed by computing a projection matrix that transforms the scalar product matrix into loadings. The projection matrix is deduced

from the combination of Equations 16 and 19, which gives

$$\mathbf{F} = \mathbf{S}_{[+]} \mathbf{Q} \mathbf{\Lambda}^{-\frac{1}{2}}. \quad (21)$$

This shows that the projection matrix is equal to $(\mathbf{Q} \mathbf{\Lambda}^{-\frac{1}{2}})$. It is used to project the scalar product matrix of each expert onto the common space. For example, the coordinates of the projections for the first expert are obtained by first computing the matrix

$$\mathbf{Q} \mathbf{\Lambda}^{-\frac{1}{2}} = \begin{bmatrix} 0.06 & -0.12 & -0.36 & -0.03 & 0.55 \\ -0.02 & -0.06 & 0.31 & 0.53 & 0.90 \\ -0.07 & -0.21 & 0.00 & -0.39 & -0.41 \\ -0.05 & 0.22 & -0.31 & 0.18 & 0.02 \\ 0.04 & 0.01 & 0.11 & 0.34 & -1.34 \\ 0.03 & 0.16 & 0.25 & -0.63 & 0.28 \end{bmatrix}. \quad (22)$$

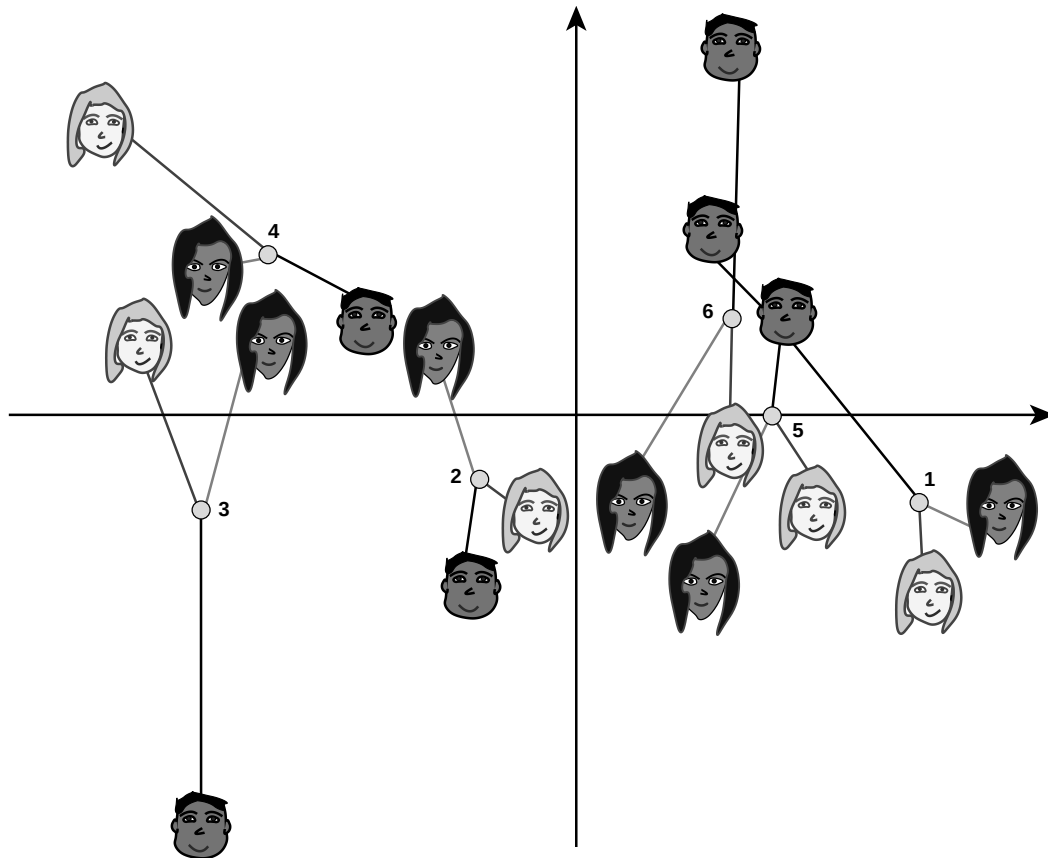
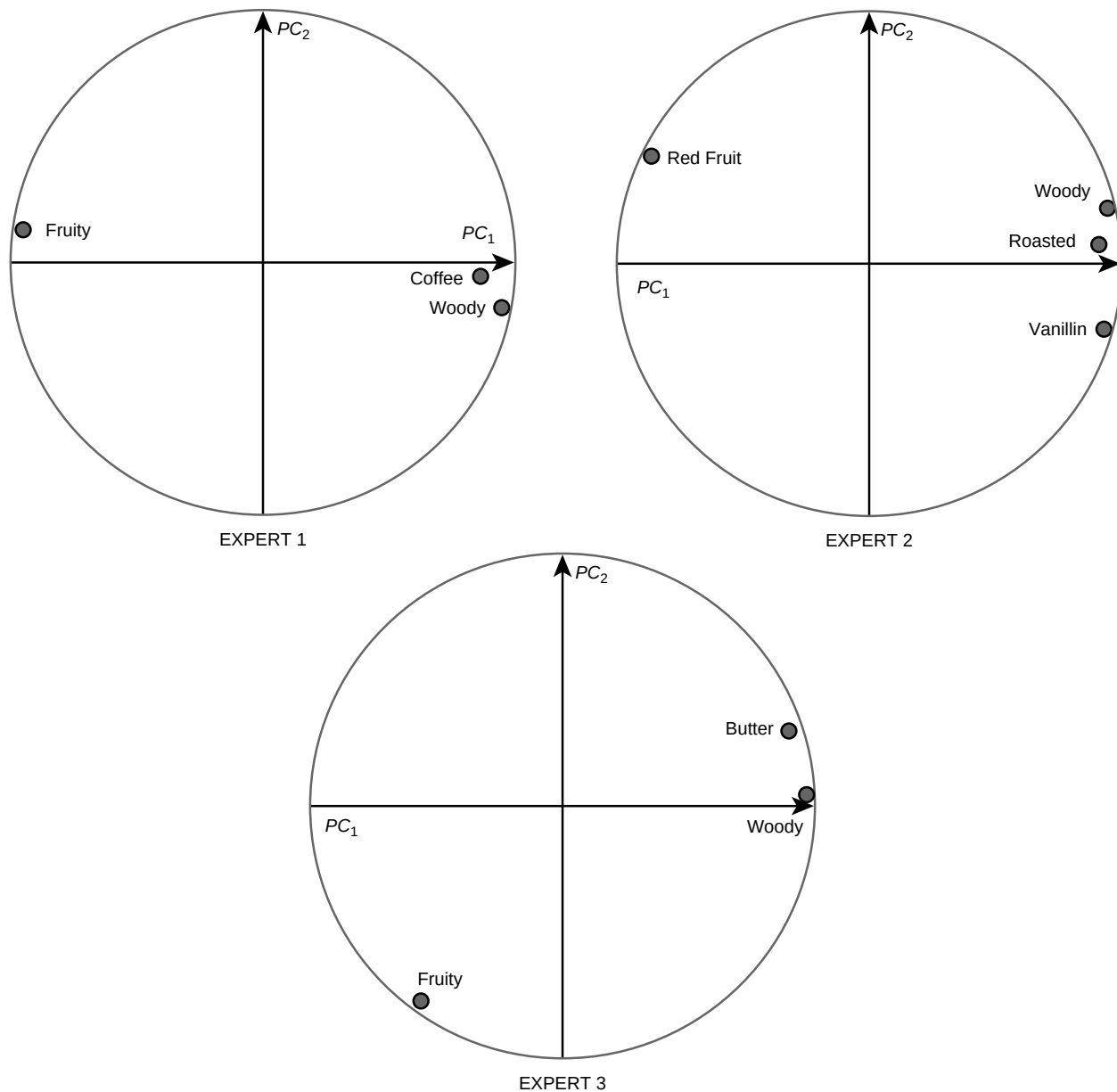


Figure 3 The Compromise: Projection of the Expert Matrices

Note: Experts are represented by their faces. A line segment links the position of the wine for a given expert to the compromise position for this wine.

Table 2 Original Variables and Compromise: Loadings (Correlation Between Variables and Components)

Axis	<i>Expert 1</i>			<i>Expert 2</i>				<i>Expert 3</i>		
	<i>fruity</i>	<i>woody</i>	<i>coffee</i>	<i>fruity</i>	<i>roasted</i>	<i>vanillin</i>	<i>woody</i>	<i>fruity</i>	<i>butter</i>	<i>woody</i>
1	-.97	.99	.92	-.90	.95	.96	.97	-.57	.94	.99
2	.19	-.13	-.04	.36	.02	-.19	.13	-.82	.22	.03
3	.00	-.00	-.36	-.22	.29	.01	-.13	.01	.25	-.10
4	.07	.08	-.07	.13	.03	.22	-.13	-.07	-.09	.04
5	.10	.00	.12	-.05	.05	.01	-.07	.00	.05	-.05

**Figure 4** Original Variables and Compromise: The Circle of Correlations for Each Expert

and then using this matrix to obtain the coordinates as

$$\mathbf{F}_{[1]} = \mathbf{S}_{[1]}(\mathbf{Q}\mathbf{A}^{-\frac{1}{2}}) \quad (23)$$

$$= \begin{bmatrix} 5.27 & -1.96 & -4.07 & -1.40 & 1.11(46) \\ -1.51 & 0.54 & 1.48 & 0.71 & -0.47(47) \\ -3.76 & 1.42 & 2.60 & 0.69 & -0.64(48) \\ -3.84 & 1.72 & 1.51 & 0.69 & 1.29(49) \\ 2.50 & -1.15 & -0.76 & -0.34 & -1.08(50) \\ 1.34 & -0.57 & -0.75 & -0.35 & -0.20 \end{bmatrix}. \quad (24)$$

The same procedure is used for experts 2 and 3 and gives

$$\mathbf{F}_{[2]} = \begin{bmatrix} 4.66 & -1.68 & 1.04 & 1.33 & 0.04 \\ -0.64 & -0.39 & 0.68 & 1.13 & 0.94 \\ -4.27 & 1.01 & -0.87 & -1.51 & 0.22 \\ -4.59 & 2.29 & -2.37 & -1.08 & -1.14 \\ 3.06 & -0.99 & 0.81 & 1.26 & 0.10 \\ 1.78 & -0.23 & 0.71 & -1.13 & -0.15 \end{bmatrix}, \quad (25)$$

and

$$\mathbf{F}_{[3]} = \begin{bmatrix} 2.96 & 1.13 & 0.45 & -0.05 & -0.68 \\ -1.10 & -1.40 & 0.06 & -0.33 & 0.35 \\ -5.17 & -7.15 & -1.78 & -0.32 & 0.05 \\ -2.30 & 0.55 & -1.37 & 1.00 & -0.12 \\ 2.65 & 2.52 & 0.75 & 0.04 & -0.24 \\ 2.96 & 4.35 & 1.89 & -0.34 & 0.65 \end{bmatrix}. \quad (26)$$

Figure 3 shows the first two principal components of the compromise space, along with the projections of wines for each expert. Note that the position of each wine in the compromise is the barycenter of the positions of this wine for the three experts. In order to make this relation clear and also to facilitate the interpretation, we have drawn lines linking the position of each wine for each expert to the compromise position. This picture confirms a conclusion obtained from the analysis of the $\mathbf{C}$ matrix: Expert 3 tends to be at variance with the other two experts.

The Original Variables and the Compromise

The analysis of the compromise reveals the structure of the set of observations, but the original data tables

were rectangular tables (i.e., each expert was using several scales to evaluate the wines). And we want to be able to relate these specific scales to the analysis of the compromise.

The original variables can be integrated into the analysis by adapting the standard approach that PCA uses to relate original variables and components, namely, computing loadings (i.e., correlation between the original variables and the factor scores). This approach is illustrated in Table 2, which gives the loadings between the original variables and the factors of the compromise. Figure 4 shows the circle of correlation obtained for each expert (these loadings could have been drawn on the same picture). Here we see, once again, that Expert 3 differs from the other experts and is mostly responsible for the second component of the compromise.

—Hervé Abdi and Dominique Valentin

See also DISTATIS; Eigendecomposition; Metric Multidimensional Scaling; Multiple Correspondence Analysis; Multiple Factor Analysis; R_v and Congruence Coefficients; Singular and Generalized Singular Value Decomposition

Further Reading

- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Chaya, C., Perez-Hugalde, C., Judez, L., Wee, C. S., & Guinard, J. X. (2003). Use of the STATIS method to analyze time-intensity profiling data. *Food Quality and Preference*, 15, 3–12.
- Coquet, R., Troxler, L., & Wipff G. (1996). The STATIS method: Characterization of conformational states of flexible molecules from molecular dynamics simulation in solution. *Journal of Molecular Graphics*, 14, 206–212.
- Escoufier, Y. (1980). L'analyse conjointe de plusieurs matrices de données. In M. Jolivet (Ed.), *Biométrie et temps* (pp. 59–76). Paris: Société Française de Biométrie.
- Kherif, F., Poline, J.-P., Mériaux, S., Benali, H., Flandin, G., & Brett, M. (2003). Group analysis in functional neuroimaging: Selecting subjects using similarity measures. *NeuroImage*, 20, 2197–2208.
- Korth, B., & Tucker, L. R. (1976). Procrustes matching by congruence coefficients. *Psychometrika*, 41, 531–535.

- Lavit, C., Escoufier Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS) method. *Computational Statistics & Data Analysis*, 18, 97–119.
- L'Hermier des Plantes, H. (1976). *Structuration des tableaux à trois indices de la statistique*. Thèse de troisième cycle, Université de Montpellier, France.
- Meyners, M., Kunert, J., & Qannari, E. M. (2000). Comparing generalized procrustes analysis and STATIS. *Food Quality and Preference*, 11, 77–83.
- Qannari, E. M., Wakeling, I., & MacFie, J. H. (1995). A hierarchy of models for analyzing sensory data. *Food Quality & Preference*, 6, 309–314.
- Robert, P. & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics*, 25, 257–265.
- Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T. Næs & E. Risvik (Eds.), *Multivariate analysis of data in sensory sciences* (pp. 229–306). New York: Elsevier.
- Stanimirova, I., Walczak, B., Massart, D. L., Simeonov, V., Sabyd, C. A., & di Crescenzo, E. (2004). STATIS, a three-way method for data analysis: Application to environmental data. *Chemometrics & Intelligent Laboratory Systems*, 73, 219–233.
- Thioulouse, J., Simier, M., & Chessel D. (2004). Simultaneous analysis of a sequence of paired ecological tables, *Ecology*, 85, 272–283.

STATISTICAL SIGNIFICANCE

Before accepting the theoretical importance or real-life impact of their research findings, psychologists have to be sure that their findings are statistically significant (i.e., that the data are not the result of happenstance). Psychologists use the *null hypothesis significance test procedure* (NHSTP) to test for statistical significance, which may be explained by exploring (a) the uncertainty inherent in empirical data, (b) the nature of inferential statistics, (c) the test statistic that represents a research outcome, and (d) the nature of the binary decision about chance effects.

Chance Effects on Empirical Data

The *substantive population* of a research effort consists of all individuals to whom the research conclusions apply. It may consist of all hyperactive boys in a

study of hyperactivity. Any of its characteristics (e.g., the mean attention span of all hyperactive boys, u) is a *parameter*. Suppose that a psychologist collects data from a randomly selected *sample* of 100 hyperactive boys. A characteristic of the sample is a *statistic* (e.g., the sample's mean attention span, $\bar{X}$).

The sample mean is unlikely to be identical to the population mean because of chance influences. For example, chance factors during data collection (a) determine who are included in the sample and (b) render some boys more attentive than usual while other boys are being distracted more than usual. Consequently, different samples of 100 hyperactive boys selected and tested in exactly the same way produce different mean attention spans.

Suppose that the psychologist selects randomly 100 hyperactive boys and assigns randomly 50 to Group 1 and 50 to Group 2. The random selection and random assignment procedures warrant the suggestion that Groups 1 and 2 are the respective samples of two substantive populations with the same mean (i.e., $u_{\text{spanI}} = u_{\text{spanII}}$). Be that as it may, the means of Groups 1 and 2 ($\bar{X}_{\text{spanI}}$ and $\bar{X}_{\text{spanII}}$, respectively) are not expected to be literally the same by virtue of happenstance.

Statistical Populations and Research Manipulation

To test whether Drug D affects the attention span of hyperactive boys, the psychologist gives Group 1 Drug D and Group 2 a placebo. The two substantive populations now become two methodologically defined statistical populations, namely, (a) hyperactive boys given Drug D and (b) hyperactive boys given a placebo. Their means are u_{DrugD} and u_{Placebo} , respectively. The means of Groups 1 and 2 are $\bar{X}_{\text{DrugD}}$ and $\bar{X}_{\text{Placebo}}$, respectively.

If Drug D is not efficacious, $u_{\text{DrugD}} = u_{\text{Placebo}}$ because the distinction between the two methodologically defined populations becomes mute. This equality implies that $\bar{X}_{\text{DrugD}} = \bar{X}_{\text{Placebo}}$. However, because of the aforementioned chance effects, $(\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}})$ is not expected to be exactly zero.

An efficacious Drug D would change the attention span, thereby leading to $u_{\text{DrugD}} \neq u_{\text{Placebo}}$. It follows that $(\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}})$ is not zero. Thus arises the following conundrum: A nonzero $(\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}})$ is expected

regardless of the efficacy of Drug *D*. Psychologists use statistical significance to handle the dilemma.

Inferential Statistics

Psychologists use *NHSTP* to learn something about population parameters (e.g., $u_{\text{DrugD}} - u_{\text{Placebo}}$) on the basis of the statistical significance of their corresponding sample statistics ($\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}}$). This is achieved by using the standardized form of an appropriate theoretical distribution to make a binary decision regarding chance effects on the data in probabilistic terms.

The Theoretical Distribution

The statistical question is whether or not the observed $\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}}$ warrants the conclusion that $u_{\text{DrugD}} \neq u_{\text{Placebo}}$. Psychologists answer this question by appealing to an appropriate *random sampling distribution of differences* that describes the probabilities of all possible values of $\bar{X}_{\text{DrugD}} - \bar{X}_{\text{Placebo}}$.

A random sampling distribution of differences is a theoretical probability density function defined by its mean (called *mean difference*) and standard deviation (called *standard error*). The two parameters are mathematically related to the means and standard deviations of the two underlying statistical populations, respectively. Specifically, the mean difference ($u_{\text{DrugD}} - u_{\text{Placebo}}$) equals ($u_{\text{DrugD}} - u_{\text{Placebo}}$), and the *standard error of differences* is a function of the two population standard deviations. The latter function depends, moreover, on sample size.

The Test Statistic

Consequently, a different random sampling distribution is implicated when there is a change in the sample sizes. This gives rise to an infinite number of probability density functions. The situation is made less daunting by standardizing the random sampling distribution, using its standard error as the unit of measure. The result is a distribution of the *t* statistic for every combination of two sample sizes. The *t* statistic is given by Equation 1:

$$t = \frac{\text{Sample statistic} - \text{Population parameter}}{\text{Standard error of the random sampling distribution}} \quad (1)$$

Specifically, for the hyperactivity example,

$$t = \frac{(\bar{X}_1 - \bar{X}_2)(u_1 - u_2)}{\text{Standard error of differences}}.$$

The *t* statistic is an example of a *test statistic*, namely, the statistic used in making the statistical decision about chance.

A Binary Decision Task

The rationale of tests of statistical significance may now be illustrated with a binary decision task. The decision maker is confronted with two sets of 50 scores each. Set *Null* is depicted in Table 1. The probability of every possible score-value is also shown. Nothing is known about the other set, Set *Alternative*, except that (a) its scores overlap with those of the Set *Null* to an unknown extent, (b) a score, *X*, is selected on any occasion from either set with an unknown probability, and (c) all scores within the chosen set have an equal chance of being selected. The decision maker has to indicate from which set the score is chosen.

The decision is made solely on the basis of Set *Null* because nothing is known about Set *Alternative*. A rational rule is to consider Set *Null* too unlikely as the source of Score *X* if its associated probability (i.e., the probability of obtaining a value that is equal to, or more extreme than, *X*) is smaller than a predetermined value (e.g., 0.04). Specifically, the decision maker rejects Set *Null* as the source when 12 or 2 is chosen because the associated

Table 1 The Simple Frequency Distribution of 25 Scores in Set *Null*

Score	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	1	1	3	4	5	7	8	7	5	4	3	1	1
Probability	.02	.02	.06	.08	.1	.14	.16	.14	.1	.08	.06	.02	.02

probability of either of them is 0.04. The value 12 or 2 is a criterion value for the decision (hence called a *critical value*).

The Statistical Hypotheses

A test of statistical significance is a binary decision task. Analogous to Score X in the binary decision task is the result of a research effort, such as $(\bar{X}_{DrugD} - \bar{X}_{Placebo})$. Analogous to Set *Null* is the appropriate chance-dependent t distribution chosen with reference to the degrees of freedom (df ; viz. $n_1 + n_2 - 2$ in the present example). *NHSTP* begins with two statistical hypotheses, H_0 and H_1 .

H_0 : An Implication of Chance Effects

Accepting the inevitability of chance effects, the psychologist allows for a nonzero $(\bar{X}_{DrugD} - \bar{X}_{Placebo})$ even though $(u_{DrugD} = u_{Placebo})$ is envisaged when Drug D is inefficacious. Encapsulating this methodological frame of reference are the conditional propositions [CP1-a] and [CP1-b].

[CP1-a]: If the chance effects assumption is true, then $H_0: u_{DrugD} = u_{Placebo}$.

[CP1-b]: If H_0 is true, then the random sampling distribution of differences has a mean difference of zero (i.e., $u_{DrugD} - u_{Placebo} = 0$).

[CP1-a] shows that the *statistical null hypothesis* (H_0) is an implication of the chance effects assumption. [CP1-b] makes explicit that H_0 stipulates the to-be-used random sampling distribution of differences linking $(\bar{X}_{DrugD} - \bar{X}_{Placebo})$ to $(u_{DrugD} - u_{Placebo})$. The statistical decision about chance effects is made with the standardized form of this random sampling distribution (see “The Test Statistic” section above).

The Complement of Chance Effects, H_1

The complement of the chance effects assumption is the stance that influences

other than happenstance determine $(\bar{X}_{DrugD} - \bar{X}_{Placebo})$. Its implication is the statistical alternative hypothesis (H_1), as may be seen from [CP2-a] or [CP2-b] or [CP2-c].

[CP2-a]: If there is more than chance effects, then $H_1: u_{DrugD} \neq u_{Placebo}$.

[CP2-b]: If there is more than chance effects, then $H_1: u_{DrugD} > u_{Placebo}$.

[CP2-c]: If there is more than chance effects, then $H_1: u_{DrugD} < u_{Placebo}$.

The consequent of [CP2-a] is a two-tailed H_1 , in which the direction of $(u_{DrugD} - u_{Placebo})$ is not specified. The consequent of [CP2-b] or [CP2-c] is a one-tailed H_1 , in which the direction of $(u_{DrugD} - u_{Placebo})$ is specified (see Table 2 for the exact relationship).

Researchers carry out research such as this because they do not know the magnitude of $(u_{DrugD} - u_{Placebo})$. This situation is analogous to knowing nothing about Set *Alternative*. That is, H_1 does not stipulate a usable random sampling distribution of differences because $(u_{DrugD} - u_{Placebo})$ is not known. It is for this reason that H_1 plays no part in the statistical-significance decision.

Some psychologists suggest that the ill-defined sampling distribution implied by H_1 serves to inform the psychologist of the probability of obtaining statistical significance (viz., *statistical power*). Be that as it may, by virtue of the theoretical implication or pragmatic objective of the research, H_1 informs (a) whether the significance test is directional and (b) the direction of the difference if it is a directional test. At the same time, there are reasons to question the validity of statistical power.

Table 2 The Pairing of H_0 and H_1

	If the chance effects assumption is true, then	If there is more than chance effects, then
Nondirectional test	$H_0: u_{DrugD} = u_{Placebo}$	$H_1: u_{DrugD} \neq u_{Placebo}$
Directional test	$H_0: u_{DrugD} \leq u_{Placebo}$ $H_0: u_{DrugD} \geq u_{Placebo}$	$H_1: u_{DrugD} > u_{Placebo}$ $H_1: u_{DrugD} < u_{Placebo}$

The Statistical Decision About Chance Effects

The psychologist adopts a predetermined criterion (called α , which is by convention 0.05) for rejecting the chance-effects explanation of the data. The α level is also known as the *level of significance*. H_0 is rejected if the computed t equals to, or exceeds, a critical t (viz., a t whose associated probability is 0.05). The result is deemed statistically significant at the 0.05 level in such an event. The critical value of t is analogous to the critical 12 or 2 in the binary decision task.

Two Types of Errors

It is clear that a statistically significant result is one that is deemed too unlikely to be due to happenstance. The “too unlikely” characterization sets in high relief that the binary decision is liable to be mistaken. There are two types of errors. A Type I error is committed if H_0 is rejected when H_0 is true. A Type II error is made if H_0 is not rejected when H_1 is true.

Although psychologists do not know the probability of the Type II error (because the implication of H_1 is unknown), they can (and do) set the probability of the Type I error (called α). To say that a research result is significant at the 0.05 level is to say that it might be brought about by chance fewer than five times out of 100 *in the long run*. It says, at the same time, that the psychologist would be wrong five times out of 100 in the long run if H_0 is rejected.

Summary and Conclusions

Statistically significant results are outcomes that are deemed too unlikely to be due to chance event. The decision is based on a well-defined random sampling distribution. The practical value or theoretical importance of the result is not informed by statistical significance because statistical decision, value judgment, and theoretical analysis belong to different domains.

—Siu L. Chow

Further Reading

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Chow, S. L. (1991). Some reservations about statistical power. *American Psychologist*, 46, 1088–1089.
- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

STEM-AND-LEAF DISPLAY

The stem-and-leaf display is an exploratory data analysis technique developed by John Tukey to summarize graphically the characteristics of a distribution. It is especially easy to produce by hand, and it is effective for examining distributions for small sample sizes. (All major statistical packages provide point-and-click routines that can be used readily to examine the distributions in the case of large sample sizes.) At the top of the next page are side-by-side stem-and-leaf displays for 100 randomly selected IQ scores and the subset of all scores greater than 99.

To construct a stem-and-leaf display, we place the data in ascending order and break each score into two parts, a stem (the leading digit[s], which contains the most salient information) and a leaf (the trailing digit[s], which contains less salient information). Two common line depths (i.e., “stem widths”) are created by forming stems that include the digits 0–4 and 5–9 (see Random IQ display) or by forming stems that include the digits 0–1, 2–3, 4–5, 6–7, and 8–9 (see IQ>99 display).

In the stem-and-leaf display for the variable Random IQ, the first line (1 6 4) indicates there is one score in the line (“line depth”), the IQ score 64, which is partitioned into a stem (6) and a leaf (4). For scores below the median, line depths are calculated by counting the total number of scores from the bottom (lowest score) of the distribution up to, and including, the scores in that line. For scores above the median, line

Stem-and-leaf of Random IQ N = 100

Leaf Unit = 1.0

Depth Stem Leaf

1 6 4

1 6

3 7 23

6 7 678

7 8 3

13 8 556789

22 9 011133444

31 9 667788999

(20) 10 0001112223333344444

49 10 55555677788888899

32 11 001111122333

20 11 5566788899

10 12 012244

4 12 58

2 13 01

Stem-and-leaf of IQ > 99 N = 69

Leaf Unit = 1.0

6 10 000111

15 10 222333333

25 10 4444455555

29 10 6777

(8) 10 88888899

32 11 00111111

25 11 22333

20 11 55

18 11 667

15 11 88899

10 12 01

8 12 22

6 12 445

3 12

3 12 8

2 13 01

depths are calculated by counting the total number of scores from the top (highest score) of the distribution down to, and including, the scores in that line. For the line that contains the median (IQ score 104), the depth of the line is reported in parentheses (20). Note that the center three depths (31, 20, 49) add to $N = 100$ since they account for all the scores within the distribution. Note also that some computer programs report leaf frequency instead of depths for all lines.

We can locate the center of the distribution (i.e., find the median) by eye or by counting in from either end until half of the observations are counted. We can examine the range of the data by locating the minimum and maximum values. We can locate the quartiles of the distribution either by eye or by counting. Turned on its side, the stem-and-leaf display forms a “digital histogram” that can be used to examine distribution shape, locate peaks and gaps, and identify unusual observations.

For the variable Random IQ, the distribution is approximately normal. There is a gap in the distribution and a possible outlier (64). For the variable IQ>99, the distribution is positively skewed, with a gap and two possible outliers. If the line depths (stem widths) were identical in the two displays, the

stem-and-leaf display for variable IQ>99 would be identical to the bottom half of the stem-and-leaf display for variable Random IQ.

—Ward Rodriguez

See also Exploratory Data Analysis; Frequency Distribution

Further Reading

- Emerson, J. D., & Hoaglin, D. C. (1983). Stem-and-leaf displays. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 7–32). New York: Wiley.
- Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (pp. 16–19). Beverly Hills, CA: Sage.
- Tukey, J. W. (1977). *Exploratory data analysis*. Menlo Park, CA: Addison-Wesley. [See Chapter 1, “Scratching down numbers (stem-and-leaf)”]
- Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury Press. [See Chapter 1, “Stem-and-leaf displays”]

STRATIFIED RANDOM SAMPLING

Stratified random sampling is a sampling technique in which the population is divided into groups called

strata. The idea behind stratified sampling is that the groupings are made so that the population units within a group are similar. The variance of the sample estimator is the weighted sum of the within-stratum variances. Because the groupings have been made so that units within a stratum are similar, strata should be less variable than the population as a whole. The variance of the sample estimator is a measure of how different the sample estimates are from all the theoretically possible samples that could be taken (given the sample design) and is therefore a measure of precision of the sample. A sample design with low variance is more desirable than a design with high variance. With a stratified random sample in which each stratum is surveyed using simple random sampling, provided the population has been divided into homogeneous strata, the weighted sum of the within-stratum variances will be less than the variance from simple random sampling with no stratification. Stratified random sampling is therefore considered a more precise sampling technique than simple random sampling.

Example

The data set in Table 1 is the number of possums caught in traps in a study area in New Zealand. The area was divided into three strata based on a map of the area. By looking at the map and studying the mapped hill slope and aspect, the researcher divided the study area into three contiguous strata roughly corresponding to hillside easterly aspect, valley floor, and hillside westerly aspect.

The mean for stratum h is calculated as

$$\bar{X}_h = \sum_{i=1}^h \frac{X_{hi}}{n_h},$$

where n_h is the number of units selected from the h th stratum and x_{ih} is the value of the i th sample unit in the h th stratum.

The sample mean from stratified sampling is the weighted sum of the stratum means, calculated as

Table 1 Different Strata for Random Sampling

Stratum 1	Stratum 2	Stratum 3
2	2	0
0	4	1
1	2	2
4	0	3
1	2	0
0	0	0
4	4	1
2	0	2
0	6	4
6	1	1

$$\bar{X}_{st} = \frac{1}{N} \sum_{h=1}^H N_h \bar{X}_h,$$

where N is the size of the population and N_h is the size of the h th stratum.

The estimated variance of the sample mean is calculated as

$$\text{var}(\bar{X}_{st}) = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h},$$

where s_h^2 is the sample variance for the h th stratum.

In this example, the stratum means were $\bar{X}_1 = 2.00$, $\bar{X}_2 = 3.50$, and $\bar{X}_3 = 1.40$, and $N_1 = 200$ hectares, $N_2 = 150$ hectares, and $N_3 = 450$ hectares. The data show some differences in possum numbers among the strata, with half again as many possums caught in stratum 2 as in stratum 3. The sample variances were $s_1^2 = 3.80$, $s_2^2 = 5.65$, and $s_3^2 = 1.64$. With these stratum statistics, the stratified sample mean and estimated variance of the sample mean are calculated as

$$\begin{aligned} \bar{X}_{st} &= \frac{1}{800} (200 \times 2.00 + 150 \times 3.50 + 450 \times 1.40) \\ &= 1.94 \end{aligned}$$

and

$$\begin{aligned}\widehat{\text{var}}(\bar{X}_{st}) &= \frac{1}{800^2} \left(200^2 \times \left(1 - \frac{10}{200} \right) \frac{3.80}{10} \right. \\ &\quad \left. + 150^2 \times \left(1 - \frac{10}{150} \right) \frac{5.65}{10} \right. \\ &\quad \left. + 450^2 \times \left(1 - \frac{10}{450} \right) \frac{1.64}{10} \right) \\ &= 0.09.\end{aligned}$$

If stratified sampling had not been used and instead the population had been surveyed by simple random sampling and the resultant sample had the same 30 values, the sample mean would have been 2.30 and the estimated variance of the sample mean 0.15. The stratified sample has a lower variance than the equivalent simple random sample and is estimated to be more precise.

Allocation of Survey Effort Among Strata

Further gains in sample precision (measured by the reduction in estimated variance of the sample mean) could be made by changing how much sample effort is expended within each stratum. In the possum example, stratum 3 had the lowest sample variance. Reducing the sample size within this stratum could mean that the strata with the higher variances could have a higher sample size. This should reduce the variance of the stratified sample. Conversely, stratum 3 is the largest-size stratum, and a large sample within this stratum could be warranted on this ground.

Optimal allocation of survey effort among strata to achieve the lowest variance for a given sample size incorporates both stratum variance and size. Differential sample costs among strata can also be included as a parameter in optimal allocation:

$$n_h = n \times \left(\frac{N_h \sigma_h}{\sqrt{c_h}} / \sum_{h=1}^H \frac{N_h \sigma_h}{\sqrt{c_h}} \right),$$

where c_h is the cost of sampling a unit in the h th stratum.

The result of optimal allocation is that strata that are variable, large, and proportionally cheap to survey are allocated larger sample sizes compared with strata that have low variance, are small, and are expensive to survey.

With the example data set, assuming the cost of surveying a unit was the same for all strata and that the within-stratum variances were $\sigma_1^2 = 3.80$, $\sigma_2^2 = 5.65$, and $\sigma_3^2 = 1.64$, the optimal allocation for $n = 30$ among the three strata can be estimated. Optimal allocation when the costs of sampling are assumed to be equal among the strata is called *Neyman allocation*.

For stratum 1,

$$\begin{aligned}n_1 &= 30 \times \left(\frac{200\sqrt{3.80}}{200\sqrt{3.80} + 150\sqrt{5.65} + 450\sqrt{1.64}} \right) \\ &= 8.84 \\ &= 9 \text{ (rounded)}.\end{aligned}$$

For stratum 2,

$$\begin{aligned}n_2 &= 30 \times \left(\frac{150\sqrt{5.65}}{200\sqrt{3.80} + 150\sqrt{5.65} + 450\sqrt{1.64}} \right) \\ &= 8.09 \\ &= 8 \text{ (rounded)}.\end{aligned}$$

For stratum 3,

$$\begin{aligned}n_3 &= 30 \times \left(\frac{450\sqrt{1.64}}{200\sqrt{3.80} + 150\sqrt{5.65} + 450\sqrt{1.64}} \right) \\ &= 13.07 \\ &= 13 \text{ (rounded)}.\end{aligned}$$

Optimal allocation would mean that stratum 1 has a sample size of 9, stratum 2 has a sample size of 8, and stratum 3 has a sample size of 13.

A simpler method of allocating survey effort among strata is to use proportional allocation, in which effort is allocated according to the relative size of the strata:

$$n_h = n \times \left(\frac{N_h}{\sum_{h=1}^H N_h} \right)$$

If sample effort had been allocated only on the proportional sizes of the strata, then stratum 1 would have been allocated a sample size of $n_1 = 7$, stratum 2 a sample size of $n_2 = 6$, and stratum 3 a sample size of $n_3 = 17$:

$$\begin{aligned} n_1 &= 30 \times \left(\frac{200}{800} \right) \\ &= 7.50 \\ &= 7 \text{ (rounded),} \\ n_2 &= 30 \times \left(\frac{150}{800} \right) \\ &= 5.63 \\ &= 6 \text{ (rounded),} \end{aligned}$$

and

$$\begin{aligned} n_3 &= 30 \times \left(\frac{450}{800} \right) \\ &= 16.88 \\ &= 17 \text{ (rounded).} \end{aligned}$$

The sample size for stratum 1 was rounded to $n_1 = 7$ rather than 8 to ensure the total sample size was 30 (if it were rounded to 8, then the total sample size would have been 31).

Proportional allocation is simpler in that estimates of within-stratum variances and sample costs are not needed. However, assuming the estimates of within-stratum variances and sample costs are accurate, optimal allocation will result in smaller variance of the stratified sample estimate than proportional allocation will.

With optimal allocation, in the example data set, stratum 3 was allocated the largest sample size. Stratum 3 was the largest stratum ($N_3 = 450$, compared with $N_1 = 200$ and $N_2 = 150$) but was also the least variable stratum ($s_3^2 = 1.64$, compared with $s_1^2 = 3.8$ and $s_2^2 = 5.65$). Optimal allocation resulted in a smaller sample size for stratum 3 than the proportional allocation sample size, reflecting the fact that optimal

allocation incorporates stratum variance (and cost) along with stratum size in allocating survey effort.

Advantages of Stratified Sampling

Other than improved survey precision, stratified sampling offers the advantage that it is often logistically easier to divide a population into strata and then plan to survey within each stratum. For example, in a field survey for a particular species of plant, if the study area is very large, it may be convenient to divide the study area into strata. Separate survey teams can then survey within each stratum. A further advantage of stratified sampling is that surveying each stratum separately allows separate estimates of the mean or total to be calculated for each stratum. In the field study example, it may be very useful to know the estimated total number of plants in the separate strata so that stratum-specific decisions about how to manage the plants can be made.

Defining Strata

Stratum boundaries can be defined on the basis of any criteria and any number of criteria. The idea behind stratified sampling is that units within strata should be as similar as possible, so it is sensible to have stratum boundaries that relate in some way to the experimental unit. In the possum survey, the strata were defined on geographical information and with the knowledge that possum densities vary among hillsides with an easterly aspect, valley floors, and hillsides with a westerly aspect. Stratified random sampling may be used to estimate how much money commercial businesses spend on staff training (per staff member), and in this application, it would be sensible to stratify on business size. Large and small businesses are likely to spend different amounts per staff member, whereas it seems reasonable that most large businesses spend roughly similar amounts.

—Jennifer Ann Brown

Further Reading

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Thompson, S. K. (1992). *Sampling*. New York: Wiley.

Stratified sampling (and other sampling methods) simulation:
<http://www.soc.surrey.ac.uk/samp/>

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Lakhani, R. (2004). Occupational health of women construction workers in the unorganized sector. *Journal of Health Management*, 6(2), 187–200.

This study assessed the occupational health status of women workers in the construction industry by evaluating incidences of occupational health disorders. In all, 1,052 workers were selected by **stratified random sampling**; medically examined; and subjected to relevant interviews, examinations, and investigations. More than three fourths of the women and almost all the men reported working 10 to 12 hours daily. A majority of the women reported headaches and backaches, as well as pain in the limbs. As much as 56% of women and 16% of men reported injuries resulting in work loss and had no social security or other workers' benefits. Respiratory, eye, and skin disorders and noise-induced hearing loss were found to be prevalent among workers exposed to hazards like dust, noise, heat, cold, nonionizing radiation, dry cement, glass, adhesives, tar, and paint. About 76% of the women reported gender-specific work stress factors, such as sex discrimination and balancing work and family demands, above and beyond the impact of general job stressors such as job overload and skill underutilization.

STRONG INTEREST INVENTORY

The Strong Interest Inventory (SII; publisher: Consulting Psychologists Press, www.cpp.com) first was published by Edward K. Strong, Jr., in 1927. Over the decades, this measure of vocational interests has been updated and expanded appreciably. By virtue of its long history, the SII is one of the most well-documented instruments available for use by assessment experts. Clients find SII scores useful for making vocational and educational decisions, confirming occupational choices, suggesting new directions within a career, understanding job dissatisfaction,

and developing plans for retirement. The questionnaire, which takes about 30 minutes to complete, includes 291 items. The profile reports five types of scales—General Occupational Themes, Basic Interest Scales, Occupational Scales, Personal Style Scales, and Administrative Indexes.

The General Occupational Themes are designed to measure John Holland's vocational types—Realistic (building, working with machines), Investigative (researching, analyzing data), Artistic (creating, writing, performing), Social (helping others, teaching), Enterprising (selling, leading), and Conventional (paying attention to detail, organizing). The Basic Interest Scales (BIS) cover 30 content areas, each of which is clustered according to its strongest relation with Holland's vocational types. For example, Realistic BIS includes mechanics and construction, Investigative BIS includes medical science, Artistic BIS includes visual arts and design, Social BIS includes counseling and helping, Enterprising BIS includes entrepreneurship, and Conventional BIS includes office management. The General Occupational Themes and the BIS are normed on a large sample of employed women and men.

The Occupational Scales measure the interests of 122 occupations, drawn from a wide array of professional, technical, and nonprofessional jobs, with separate sex-normed scales reported for each occupation (i.e., a total of 244 Occupational Scales). The scores for these scales reflect the similarity of a client's interests to those of people in the occupation.

The Personal Style Scales measure an individual's work and learning style. Work Style is designed to distinguish between people who like working with others (e.g., school counselors) and those who prefer working with data, things, and ideas (e.g., mathematicians). Learning Environment identifies people who prefer academic environments (e.g., sociologists) versus those who prefer hands-on environments (e.g., auto mechanics). Leadership Style differentiates people who like to take charge (e.g., school administrators) and those who prefer to follow (e.g., production workers). Risk Taking identifies people who like risky activities (e.g., firefighters) and those who are cautious (e.g., librarians). Team Orientation contrasts

those who prefer team-based activities (e.g., nursing home administrators) and those who prefer working alone (e.g., medical illustrators).

Studies have demonstrated the reliability of SII scale scores over short and relatively long periods of time. An extensive body of literature also reports evidence of validity for SII scores for use with clients who vary in age, ability, ethnicity, and interest orientation.

—Jo-Ida C. Hansen

Further Reading

- Donnay, D., Morris, M., Schaubhut, N., & Thompson, R. (2005). *Strong Interest Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Hansen, J. C. (1992). *User's guide for the Strong Interest Inventory*. Stanford, CA: Stanford University Press.

STROOP COLOR AND WORD TEST

The Stroop Color and Word Test developed from the observation by early experimental psychologists that the naming of color hues is always slower than the reading of color names in literate adults. The earliest published report of this phenomenon was offered by Cattell in 1886. J. Ridley Stroop suggested in 1935 that the difference in color naming and word reading was due to colors' being associated with a variety of behavioral responses while words were associated with only one behavioral response: reading. In order to further study the relationship between color naming and word reading, Stroop devised the test that has come to be called the Stroop Color and Word Test.

The earliest version of the Stroop consisted of the words *red, green, brown, blue, and purple* printed on a page consisting of ten rows and ten columns. Each word was printed in colored ink, but never in the color it represented by the word (e.g., *red* could be printed in blue ink but never in red ink). Another page of the test consisted of colored ink printed as small rectangles. The final page consisted of the color words listed above, this time printed in black ink.

Since Stroop's original studies, several hundred studies have been published on the Stroop test. The Stroop has attracted such attention because of its high reliability in identifying individual differences and because of its somewhat paradoxical nature. The research has examined the use of the Stroop in cognitive and personality research, in experimental psychopathology, and in the diagnosis and understanding of organic brain dysfunction.

Neuropsychological studies have suggested that Stroop interference occurs not at the response stage or in the confusion of the participant but as a result of interference in verbal processing. The Stroop stimuli appear to activate an automatic verbal processing response that interferes with the consciously instructed color naming. The participant completes the task either by completing both responses sequentially (reading the word, followed by naming the color) or by suppressing the automatic, word-reading response through volitional control.

Current versions of the Stroop test consist of three pages and generally use only three colors. Page 1 consists only of color word names (e.g., *red, green, and blue*) while page 2 uses XXXX's printed in red, green, and blue ink. The third page (interference) consists of the color words on page 1 printed in nonmatching ink (e.g., the word *red* printed in blue ink). The participant must name the color of the ink rather than the word. The Stroop may be scored as the total number of items finished in a set time (usually between 30 and 60 seconds) or the time to complete an entire page. The former version has the advantage of limiting the time the test takes without apparently limiting the information that is gathered. It is used widely as a neuropsychological instrument, primarily as a measure of possible reading disorders (when there is no interference effect) and as a measure of complex executive functions (when the interference effect is unusually large).

—Charles Golden

Further Reading

- Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, 11, 63–65.

- Dyer, F. N. (1973). The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. *Memory & Cognition*, 1(2), 106–120.
- Golden, C. J. (1976). The diagnosis of brain damage by the Stroop test. *Journal of Clinical Psychology*, 32, 654–658.
- Jensen, A. R., & Rohwer, W. D., Jr. (1966). The Stroop color-word test: A review. *Acta Psychologica*, 25, 36–93.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- Wheeler, D. D. (1977). Locus of interference on the Stroop test. *Perceptual and Motor Skills*, 45, 263–266.

STRUCTURAL EQUATION MODELING

Structural equation modeling (SEM) is a general term that describes a large number of statistical models used to evaluate the consistency of substantive theories with empirical data. It represents an extension of general linear modeling procedures such as analysis of variance and multiple regression. In addition, SEM can be used to study the relationships among latent constructs that are indicated by multiple measures and is applicable to experimental or nonexperimental data and to cross-sectional or longitudinal data.

SEM has a number of synonyms or special cases in the literature, including *path analysis*, *causal modeling*, and *covariance structure analysis*, or some variations of these terms. Path analysis is an extension of multiple regression in that various multiple regression models or equations can be estimated simultaneously; it provides a more effective and direct way of modeling mediation effects. Path analysis can be considered an early form and a special case of SEM in which structural relations among only observed variables are modeled. Structural relations are hypotheses about how independent variables affect dependent variables. Hence, the earlier path analysis or the later SEM is sometimes referred to as causal modeling. Because analyzing interrelations among variables is a major part of SEM and these interrelations are supposed to generate the observed covariance or correlation patterns among the variables, SEM is also sometimes called covariance structure analysis.

The measurement of latent variables originated from psychometric theories. Unobserved latent variables cannot be measured directly but are indicated or inferred by responses to a number of observable variables (indicators). Latent constructs such as intelligence or reading ability are often gauged by responses to a battery of items that are designed to tap those constructs. Responses of a participant are supposed to reflect where the participant stands on the scale of the latent variable. Factor analyses have been widely used to examine the number of latent constructs underlying the observed responses and to evaluate the adequacy of individual items or variables as indicators for the latent constructs they are supposed to measure. The merge of confirmatory factor analysis models (sometimes called *measurement models*) with structural path models on the latent constructs became a general SEM framework in analyzing covariance structures. Today's advancement of SEM includes the modeling of mean structures in addition to covariance structures, the modeling of growth or changes over time (growth models), and the modeling of data having nesting structures (e.g., students are nested within classes, which in turn are nested within schools; multilevel models).

How Does SEM Work?

SEM is a complex and growing collection of techniques. This discussion will focus on only the basic logic and mechanism of the way SEM works. Interested readers should consult additional references for the specific type of model and data involved. In general, however, every SEM analysis goes through steps of model specification, data collection, model estimation, evaluation, and modification. Issues pertaining to each of these steps are discussed briefly below.

Model Specification

A sound model is theory based. According to theories, findings in the literature, knowledge in the field, or one's educated guesses, causes and effects among variables of interest can be specified. In SEM, a variable can serve both as a source (i.e., cause) variable and

a result (i.e., effect) variable in a chain of causal hypotheses. This kind of variable is often called a mediator. Suppose that social status has a direct impact on learning motivation, which in turn is supposed to affect achievement. Motivation, then, is a mediator between social status and achievement; it is the source variable for achievement and the result variable for social status. Furthermore, feedback loops among variables (e.g., achievement, in the example proposed above, can in turn affect social status) are permissible in SEM, as are reciprocal loops (e.g., learning motivation and achievement affect each other).

Models can be more easily conceptualized and communicated in graphical forms. Many specialized SEM software programs, such as LISREL and EQS, can translate graphical specification of a model into computer syntax for analysis. A directional arrow is universally used to indicate a hypothesized causal direction. For the mediation example above, the model can be expressed as social status → learning motivation → achievement. The variables to the right of the arrow heads are commonly termed *endogenous variables*, and the variables having no arrows pointing to them are called *exogenous variables*.

Given the flexibility in model specification, a variety of models can be conceived. However, not all specified models can be identified and estimated. Just as in algebra, where equations cannot have more unknowns than knowns, a basic principle of identification is that a model cannot have a larger number of unknown parameters to be estimated than the number of unique pieces of information provided by the data (variances and covariances of observed variables in conventional SEM). Another basic principle is that all latent variables must be scaled so that their values can be interpreted. These are called necessary conditions for identification, but they are not sufficient. The issue of model identification is complex. Fortunately, there are some established rules that can help researchers decide whether a particular model of interest is identified. When a model is identified, every model parameter can be uniquely estimated. When a model fails to be identified, on the other hand, either some model parameters cannot be estimated or numerous sets of parameter values can produce the same level of model

fit. In any event, results of such (unidentified) models are not interpretable, and the models require respecification (often by the addition of constraints).

Data Collection

Data collection should come, if possible, after models of interest are specified, such that sample size can be determined a priori. The sample size required to provide accurate parameter estimates and model fit information depends on the model size as well as the score scale and distribution of the measured variables. For example, larger models require larger samples to provide stable parameter estimates, and larger samples are required for categorical or nonnormally distributed variables than for continuous or normally distributed variables.

SEM is a *large-sample technique*. That is, model estimation and statistical inference or hypothesis testing regarding the specified model and individual parameters are appropriate only if sample size is not too small for the estimation method chosen. A general rule of thumb is that the minimum sample size should be no less than 200 (preferably no less than 400, especially when observed variables are not multivariate normally distributed) or 5 to 20 times the number of parameters to be estimated, whichever is larger. Larger models often contain a larger number of model parameters and hence demand larger sample sizes.

Model Estimation

A properly specified structural equation model often has some fixed parameters (e.g., missing or omitted paths have path coefficients of zero) and some free parameters to be estimated from the data. Free parameters are estimated through iterative procedures to minimize a certain discrepancy function between the observed covariance matrix (data) and the model-implied covariance matrix (model). Definitions of the discrepancy function depend on specific methods used to estimate the model parameters. A commonly used normal theory discrepancy function is derived from the *maximum likelihood method*. This estimation method assumes that the observed variables are

multivariate normally distributed or that there is no excessive kurtosis of the variables.

When this distributional assumption is violated, parameter estimates may be unbiased (if the proper covariance or correlation matrix is analyzed; i.e., Pearson for continuous variables, polychoric or polyserial correlation when categorical variables are involved), but their estimated standard errors will likely be underestimated, and the model chi-square statistic will be inflated. In other words, when distributional assumptions are violated, statistical inference may be incorrect. Other estimation methods that do not make distributional assumptions are available, but they often require unrealistically large sample sizes to work satisfactorily ($N > 1,000$). When sample size is not that large, a viable alternative is to request robust estimation from some SEM software programs (e.g., LISREL, EQS, Mplus), which provides some adjustment based on the severity of nonnormality to the chi-square statistic and standard error estimates. Statistical inference based on adjusted statistics has been found to work quite satisfactorily, provided sample size is not too small.

The estimation of a model may fail to converge, or the solutions provided may be improper. In the former case, SEM software programs generally stop the estimation process and issue an error message or warning. In the latter, parameter estimates are provided, but they are not interpretable because some estimates are out of range (e.g., correlation greater than 1, negative variance). These problems may result if a model is ill specified (e.g., the model is not identified), the data are problematic (e.g., the sample size is too small, the variables are highly correlated), or both.

Model Evaluation

Once model parameters have been estimated, one would like to make a dichotomous decision: either retain or reject the model. This is essentially a statistical hypothesis-testing problem, with the null hypothesis being that the model under consideration fits the data. The overall goodness-of-fit test statistic has a chi-square distribution (i.e., it is a chi-square test). Unfortunately, this test statistic has been found to be

extremely sensitive to sample size. That is, the model may fit the data reasonably well, but the chi-square test may reject the model because of large sample size. In reaction to this sample size sensitivity problem, a variety of alternative goodness-of-fit indices, such as *goodness of fit index*, *normed fit index*, and *comparative fit index*, have been developed to supplement the chi-square statistic. Higher values of these indices indicate better model-data fit. Values in the .90s (or more recently $\geq .95$) are generally acceptable as indications of good fit. Another popular index is the *root mean square error of approximation* (RMSEA). Unlike the other fit indices, a smaller value of RMSEA indicates better model-data fit ($\leq .05$ indicates close fit, $\leq .08$ indicates acceptable fit, $\geq .10$ indicates bad fit). SEM software programs routinely report a handful of goodness-of-fit indices. Some of these indices work better than the others under certain conditions. It is generally recommended that multiple indices be considered simultaneously when overall model fit is evaluated.

Because some solutions may be improper, it is prudent for researchers to examine individual parameter estimates as well as their estimated standard errors. Unreasonable magnitude or direction of parameter estimates or large standard error estimates are some indications of plausible improper solutions.

Model Modification

When the goodness of fit of the hypothesized model is rejected (by the chi-square model fit test and other fit indices), SEM researchers are often interested in finding an alternative model that fits the data. Post hoc modifications of the model are often based on modification indices, sometimes in conjunction with the expected parameter change statistics (which approximate the expected sizes of misspecification for individual fixed parameters). Large modification indices suggest large improvement in model fit as measured by chi-square if certain fixed parameter(s) were free to be estimated (it is less likely due to chance fluctuation if a large modification index is accompanied by a large expected parameter change value). The suggested modifications, however, may or may not be supported

on theoretical grounds. Researchers are urged not to make too many changes based on modification indices, even if such modifications seem sensible on theoretical grounds. Note that SEM takes a confirmatory approach to model testing; one does not try to find the best model or theory via data using SEM. Rather than data-driven post hoc modifications (which may be very inconsistent over repeated samples), it is often more defensible to consider multiple alternative models a priori. That is, multiple models (e.g., based on competing theories or different sides of an argument) should be specified prior to model fitting, and the best-fitting model should be selected among the alternatives. Because a more complex model, assuming it is identified, will generally produce better fit, and different models can produce the same fit, theory is imperative in model testing.

In conclusion, it is worth noting that although SEM allows the testing of causal hypotheses, a well-fitting SEM model does not and cannot prove causal relations without satisfying the necessary conditions for causal inference (e.g., time precedence, robust relationship in the presence or absence of other variables). A selected well-fitting model in SEM is like a retained null hypothesis in conventional hypothesis testing; it remains plausible among perhaps many other models that are not tested but may produce the same or better level of fit. SEM users are cautioned not to make unwarranted causal claims. Replications of findings with independent samples are recommended, especially if the models are obtained with post hoc modifications.

—Pui-Wa Lei

Further Reading

- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Schumacker, R., & Lomax, R. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Structural equation modeling, by David A. Kenny: <http://davidakenny.net/cm/causalmod.htm>

Structural equation modeling reference list (journal articles and chapters on structural equation models), by Jason Newsom: <http://www.upa.pdx.edu/IOA/newsom/semrefs.htm>

STRUCTURED CLINICAL INTERVIEW FOR *DSM-IV*

The Structured Clinical Interview for *DSM-IV* (SCID) disorders is a semistructured standardized diagnostic interview commonly used for clinical and research applications. (*DSM-IV* is the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition.) Several versions of the SCID are available, including standard face-to-face interview, computerized interviewer-entered data, and computerized patient-entered self-report. The SCID was developed at the New York Psychiatric Institute Biometrics Research Lab in a comprehensive effort to standardize *DSM-IV* diagnostic interviewing procedures including uniform phrasing of interview questions. Standardized interview procedures were greatly needed to enhance interrater *DSM* diagnostic reliability, particularly when interviewers differed significantly in education, training, and clinical experience.

When *DSM-IV* Axis I disorders are under consideration, the SCID-I interview is appropriate. The SCID-I interview is comprised of sequential modules that assess for mood disorders, psychotic symptoms, psychotic disorders, substance use disorders, anxiety, and other Axis I disorders commonly found in adults. Each module begins with screening questions. When negative responses are provided for all the screening questions, the interviewer is directed to the next module by means of a branching system. When some positive responses are provided for the screening questions, the interviewer is directed to a more comprehensive review of symptoms for that module. The SCID-II should be used when Axis II personality disorders are under consideration. Again, sequential modules are used to examine each of the *DSM-IV* Axis II personality disorders. The SCID-II can also be used for detecting the *DSM* residual category (personality disorder not otherwise specified) and some unofficial diagnostic categories (depressive personality disorder). A SCID-I or SCID-II interview typically ranges from 45 to 90 minutes, depending on the complexity of patient report, including long-standing psychiatric history.

The SCID-I and SCID-II are available in research and clinician versions. The research versions closely

follow *DSM-IV* criteria including detailed questions to ascertain presence of *DSM-IV* subtype specifiers. The research versions can be highly useful when documenting multiple *DSM* subtype specifiers, as may be encountered when interviewing a person with a complicated history of depressive or manic episodes. Although the research versions are preferred when comprehensive diagnostic information is needed to meet research protocol criteria for participant inclusion and exclusion, these versions tend to be time consuming when one is interviewing patients with intricate psychiatric histories. Rest breaks may be offered if a patient appears to be fatigued by a lengthy interview. The test developers offer a one-sided master copy of the research version and will permit modifications if the full diagnostic interview is not needed for the intended research purposes. A one-time fee with photocopying privileges is available for research funded by nonprofit organizations. A per-use fee is required when research is funded by for-profit organizations (contact Biometrics Research, 212-543-5524). Clinician versions of the SCID are streamlined to cover major DSM disorders often seen in adult clinical populations and are available from American Psychiatric Press (appi.org). Computer-assisted versions and a variety of training resources are available from Multi-Health Systems (mhs.com). On-site training and a 20-hour video training series are available from Biometrics Research. The Structured Clinical Interview for DSM-IV Dissociative Disorders (SCID-D) Revised, developed separately by Marlene Steinberg, focuses specifically on dissociation, including modules for posttraumatic dissociative symptoms and acute stress disorder (see mhs.com).

Although the SCID-I and SCID-II have been used in more than 700 published studies, it is difficult to provide validity and reliability values as study populations have varied considerably. In general, moderate-to-high validity has been demonstrated when SCID-generated diagnoses are compared across expert clinicians. Reliability estimates tend to be higher for more severe psychiatric disorders, and summaries can be found on the biometrics Web site (www.scid4.org). The SCID-I and SCID-II are readily available in English. Although significant sections of the SCID have been translated into Spanish, French,

German, Danish, and six other languages, these translations have been made on an ad hoc basis outside the Biometrics Lab. Psychometric characteristics of non-English versions are not readily available.

—Carolyn Brodbeck

Further Reading

- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV® Axis I disorders (SCID-I), clinician version, user's guide*. Washington, DC: American Psychiatric Press.
- Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. *Archives of General Psychiatry*, 49, 624–629.

Structured Clinical Interview for *DSM-IV* Web page: www.scid4.org

SUNFLOWER PLOT

A sunflower plot is an extension of the scattergram for displaying the distribution of pairs of observations. In high-density regions of a scattergram, it often becomes impossible to assess the distribution of observations because the points lie on top of each other (known as *overstriking*). Sunflower plots overcome this problem by using sunflower-like symbols to represent the density.

The two-dimensional (x, y) plane is broken up into a regular lattice of square cells (cells may also be rectangular, depending on axis scale). The number of observations in each cell is counted. A point is placed in the center of every cell that contains one or more observations. Equiangular line segments (petals) are drawn radiating from the central point. The number of line segments is equal to the number of observations. No petal is drawn for a single observation.

A drawback of sunflower plots is that in low-density regions, the location of individual observations is lost. To overcome this issue, it is recommended that for cells containing a small number of observations, the points be positioned at their actual locations.

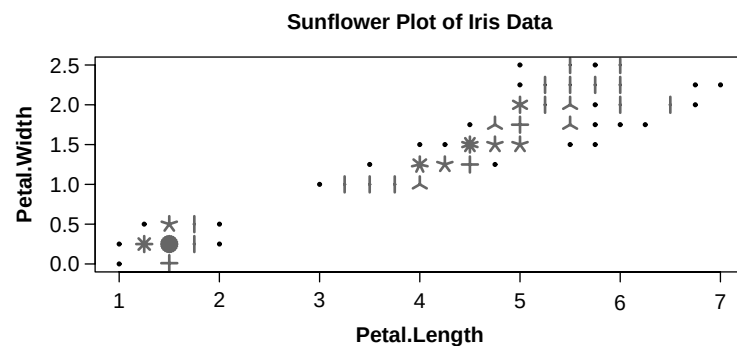


Figure 1 Sunflower Plot of Fisher's Iris Data on Petal Length and Petal Width in Centimeters

Note: The observation cells are 0.25 cm in width.

One issue with interpretation is that eyes can be drawn to meaningless artifacts as a result of the formation of neighboring sunflowers. This problem can be reduced by randomly rotating the sunflowers.

For cells of extremely high density, the petals may start to overlap and look like filled circles. In this case, it may be sensible to set the number of petals to be proportional to the number of observations (e.g., 1 petal per 10 observations).

Figure 1, created using R, is an example of a sunflower plot for Fisher's Iris data.

Many extensions of sunflower plots have been suggested. For example, the use of different symbols (e.g., circles, with size dependent on number of observations in cell) or cell shapes (e.g., hexagonal cells, which can be packed more tightly). The use of shading or colors can improve the legibility of sunflower plots and allow more information to be presented.

—Carl J. Scarrott

See also Scattergram; Smoothing

Further Reading

Cleveland, W. S., & McGill, R. (1984). The many faces of the scatterplot. *Journal of the American Statistical Association*, 79, 807–822.

Sunflower plot creation using the R software, including reference index: <http://cran.r-project.org/manuals.html>

SUPPORT VECTOR MACHINES

Support vector machines (SVM) are a system of machine learning (or classification) algorithm that constructs a classifier to assign a group label to each case on the basis of its attributes. The algorithm requires that there be two variables for each case in the data. The first is the group variable to be classified and predicted (such as disease status or treatment), and the second is the variable of attributes, which is usually multidimensional numerical data (such as amount of daily cigarette consumption or abundance of particular enzymes in the blood). Normally a

classifier is learned from the method based on a set of training cases. The classifier is then applied to an independent test set to evaluate the classification accuracy.

SVM was first developed by Vladimir Vapnik in the 1990s. In the simplest situation, binary (e.g., “disease” vs. “normal”), separable, and linear classification is considered. The method seeks a *hyperplane* that separates the two groups of data with the maximum margin, where the margin is defined as the distance between the hyperplane and the closest example in the data. The idea came originally from the Vapnik Chervonenkis theory, which shows that the generalized classification error is minimized when the margin is maximized.

Why the name *support vector machines*? The answer is that the solution of the classification hyperplane for SVM depends only on the *support vectors* that are the closest cases to the hyperplane. All the remaining cases, farther away, do not contribute to the formulation of the classification hyperplane. The optimization problem is solved through quadratic programming techniques, and algorithms are available for fast implementation of SVM.

SVM is widely applied in virtually any classification application, including writing recognition, face recognition, disease detection, and other biological problems. Two important extensions in the development of SVM made it popular and feasible

in practical applications: *kernel methods* and *soft margin*. Kernel methods are used to extend the concept of linear SVM to construct a nonlinear classifier. The idea is to map the current space to a higher-dimensional space, with the nonlinear classifier in the current space transformed to a linear one in the new, high-dimensional space. The distance structure (dot product) is simply replaced by a kernel function, and the optimization is performed similarly. Common choices of kernel functions include polynomial, radial basis, and sigmoid. Soft margin is used when the two groups of cases are not separable by any possible hyperplane. Penalties are given to “misclassified” cases in the target function, and the penalized “soft margin” is similarly optimized.

Assumptions and Applications of SVM

SVM is a distribution-free method (in contrast to methods like linear discriminant analysis). It is thus more robust to skewed or ill-behaved distributed data. The major considerations when using SVM are the selection of a proper kernel function and the parameters of penalties for soft margins. Selections of kernels and parameters are usually determined by maximizing the total accuracy in cross-validation.

Using the Computer

Since SVM is a relatively modern statistical technique, it is not implemented in most major statistical software (e.g., SAS and S-PLUS). In the following, an extension package, *e1071*, of R software is used to implement SVM, and an example of classification of car fuel efficiency is demonstrated. The data are a sub-sample of 25 cars from the “auto-mpg” data set in the UCI Machine Learning Repository. Fuel efficiency, horsepower, and weight are shown in Table 1. The goal of the classification problem is to classify inefficient and economic cars based on their attributes of horsepower and weight. In Figure 1, inefficient cars are shown on the right, and economic cars are on the left. Linear SVM is applied, and the resulting

Table 1 Fuel Efficiency, Horsepower, and Weight

Car	Efficiency	Horsepower	Weight
1	Inefficient	170	4654
2	Economic	65	2045
3	Inefficient	150	4699
4	Inefficient	225	3086
5	Inefficient	190	3850
6	Economic	70	1990
7	Inefficient	215	4312
8	Economic	62	2050
9	Economic	70	2200
10	Economic	65	2019
11	Economic	67	2145
12	Inefficient	150	4077
13	Economic	84	2370
14	Inefficient	120	3820
15	Inefficient	140	4080
16	Economic	70	2120
17	Inefficient	170	4668
18	Economic	53	1795
19	Inefficient	130	3840
20	Inefficient	150	3777
21	Inefficient	160	3609
22	Inefficient	110	3632
23	Inefficient	220	4354
24	Inefficient	110	3907
25	Inefficient	132	3455

Source: Data from the UCI Machine Learning Repository (<http://www.ics.uci.edu/~mllearn/MLRepository.html>).

Note: Inefficient: <18 mpg; economic: >32 mpg.

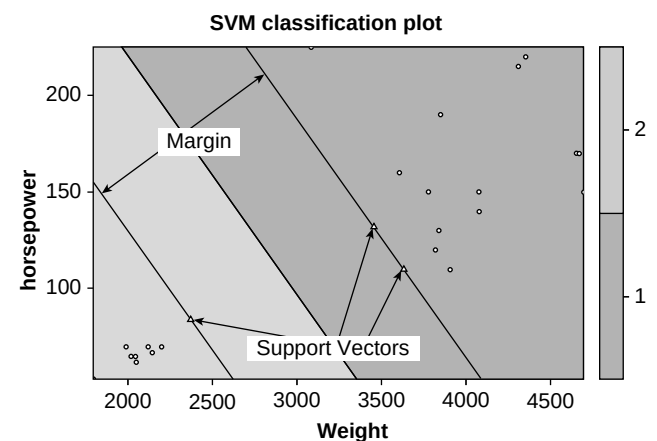


Figure 1 SVM Classification Plot for Fuel Efficiency Data

classification hyperplane, margin, and three support vectors (triangles) are indicated.

—George C. Tseng

Further Reading

- Cristianini, N., & Shawe-Taylor, J. (2001). *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press.
- Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). UCI Repository of machine learning databases, Irvine: University of California, Department of Information and Computer Science. Retrieved from <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

SUPPRESSOR VARIABLE

The suppressor phenomenon was first reported in the 1940s by P. Horst, who noted that variables that do not have a high zero order correlation with the dependent variable may contribute to an increased proportion of explained variance. He described how the selection of World War II pilots could be improved by including in the prediction equation not only a variable measuring their technical abilities but also a variable assessing their verbal ability, even though the latter variable is itself unrelated to the criterion (navigating skills). He found that the verbal ability regressor had a negative coefficient when entered into the prediction equation.

Subsequently, Quinn McNemar provided an intuitive explanation in terms of *common elements*. According to McNemar, a useful predictor has, of course, many elements in common with the dependent variable, but usually also irrelevant elements. A suppressor is a variable that has no elements in common with the dependent variable but does have irrelevant elements in common with the predictor. If the predictor and suppressor are positively correlated, then the suppressor has a negative regression weight after inclusion in the regression equation. This negative regression weight reflects the fact that the irrelevant

elements from the predictor are partialled out, which “purifies” the predictor and improves the prediction.

McNemar’s explanation proved relevant to more situations than Horst’s classical form. This is demonstrated by the examples displayed in Table 1, showing various patterns of bivariate correlations between a dependent variable Y and two independent variables X_1 and X_2 .

From Table 1, the following observations can be made:

1. The first example illustrates the *classical* suppressor condition.
2. In the second example, two independent variables have a positive zero order correlation with the dependent variable and correlate positively with each other. One of them receives a negative regression weight. This situation is referred to as *negative suppression*. Although the suppressor has relevant information in common with Y , they share fewer common elements than the common elements of irrelevant information shared by the suppressor and the other predictor.
3. The third situation involves two variables that can act as good predictors. They also share, however, information that is irrelevant to Y , but with an opposite orientation. When both variables are included in the regression equation, they suppress a part of each other’s irrelevant information. This case is called *reciprocal suppression*.
4. In all the situations, inclusion of a second predictor leads to an increase of the absolute value of the regression coefficient, as well as of the proportion of explained variance.

Table 1 Examples of Three Suppressor Situations

	1	2	3	4	5	8
<i>Situation</i>	r_{y1} (r^2_{y1})	r_{y2} (r^2_{y2})	r_{12}	$b_{y1.2}$	$b_{y2.1}$	$R^2_{y.12}$
Classical	.40 (.16)	.00 (.00)	.707	.800	-.566	.320
Negative	.50 (.25)	.10 (.01)	.710	.865	-.514	.381
Reciprocal	.50 (.25)	.30 (.09)	-.270	.627	.469	.454

On the latter characteristic, A. J. Conger based the following general definition: “A suppressor variable is defined to be a variable that increases the predictive validity of another variable (or set of variables) by its inclusion in a regression equation. This variable is a suppressor only for those variables whose regression weights are increased.” In a formula, variable X_2 is a suppressor for predictor X_1 (in relation to the dependent variable Y) if

$$\beta_1 r_{y1} > r_{y1}^2.$$

—Gerard H. Maassen

Further Reading

Maassen, G. H., & Bakker, A. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods & Research*, 30, 241–270.

SURVEY WEIGHTS

Weights are numerical values that are used in surveys to multiply by response values in order to account for missing observations. The missing data may be absent as the result of a prearranged sample design or as the result of nonresponse. In the case of sample designs, weights are used to estimate totals or means for data of interest, such as acres of corn grown or household income, based on a selected subset of the entire population. The population could be, say, all farms or all households. The subset is known as a sample. In the case of nonresponse, the weights are inflated further to account for those missing observations. Another method for accounting for nonresponse is to replace those values with data derived from other information. That process is known as *imputation*. Whether compensation for nonresponse is done by imputation, by weighting the results of a census, or by adjusting the weights for a sample, one must consider whether the missing data are different in some way from the data actually observed. Here, however, we will examine the more straightforward use of weights in surveys: sampling weights.

Data collected on a given characteristic or characteristics for a population, such as acres of corn planted on farms in Minnesota, constitute a survey. Thus a characteristic for a data element could be number of acres of corn, and a population could be all farms in Minnesota. Often we want to estimate totals or means of data elements, such as total acres of corn in a population (here, Minnesota farms), by taking a sample of members of the population and from that data, inferring an estimate for the population. The sample is selected from the population according to varying rules, depending on the type of sample. Samples may be model based or design based. Often they are design based with model-assisted inference. If they are design based, then samples are collected on the basis of the randomization principle. This practice leads to the sample weights we discuss here.

For example, the simplest design-based sample is the simple random sample (SRS). If, for example, a sample of 20 were to be drawn at random from a population with 100 members, with an equal chance of selection for each member of the population, then that sample would be an SRS. Here the sample size is $n = 20$, and the population size is $N = 100$. The probability of selection for each member of the population to become a member of the sample is $\frac{20}{100} = \frac{1}{5}$. In general, the probability of selection for an SRS is $\frac{n}{N}$. Each value we collect for a characteristic, here acres of corn, for each member of the sample can be added together for a sample total. The sample weight that we need to multiply by this sample total to obtain an estimated population total would be the inverse of the probability of selection, or $w = \frac{n}{N}$, where w is the weight. In the above example, $w = 5$.

These weights are always the inverses of the corresponding probabilities of selection. The probability of selection depends on the structure of the design-based sample and can be complex. This process may involve various stages and adjustments, but the basic fact to remember is that a sample weight for any given observation is the inverse of its probability of selection.

—James Randolph Knaub, Jr.

Further Reading

- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). *Sample survey methods and theory* (2 vols.). New York: Wiley. (Original work published 1953)
- Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Journal of Official Statistics online: <http://www.jos.nu> (Click on "Search," and then search keyword "weight." A number of articles will be found on topics such as noncoverage, raking, and other advanced topics.)

SURVIVAL ANALYSIS

Survival analysis is a well-developed branch of statistics that concerns methods for the analysis of time-to-event data. Such data arise in a number of scientific fields, including medicine, biology, public health, epidemiology, engineering, economics, and demography, among others. Time-to-event data, sometimes also referred to as failure time data (in which case the event is regarded as a "failure"), have two unique features. First, the response of interest, time, is always nonnegative. Second, and more important, these data are often *censored* or *truncated* or both, making the survival analysis unique among most other statistical methods.

Survival analysis often models survival function and hazard function instead of the probability density function or cumulative distribution function. Throughout this discussion, we will assume that the time to event, denoted by T , is a nonnegative continuous random variable with a *probability density function* $f(t)$ and that $F(t) = P(T \leq t) = \int_0^t f(u) du$ is the corresponding *cumulative distribution function*. The *survival function* and the *hazard function* of T are then defined as $S(t) = P(T > t) = 1 - F(t)$ and $\lambda(t) = \lim_{\Delta t \rightarrow 0} P(t < T \leq t + \Delta t | T > t) / \Delta t$, respectively. It can be easily shown that $\lambda(t) = f(t) / S(t)$ and $S(t) = \exp\{-\Lambda(t)\}$ where $\Lambda(t) = \int_0^t \lambda(u) du$ is termed the *cumulative hazard function*. We see that the survival function gives the probability that an individual "survives" up to time t , while the hazard function represents the "instantaneous"

probability that the "failure" will occur in the next moment, given that the individual has survived up to time t . Both the survival function and the hazard function can be used to characterize the stochastic behavior of the random variable T .

Censoring and Truncation

Time-to-event data present themselves in unique ways that create special obstacles in analyzing such data. The most important one is known as *censoring*. Loosely speaking, censoring means that the data are incomplete. Although there exist many censoring mechanisms, we shall mainly focus on the two most frequently encountered types of censoring: *right censoring* and *interval censoring*.

Table 1 presents a typical right-censored data set resulting from a clinical trial. The purpose of this clinical trial was to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML). In all, 23 patients were treated by chemotherapy, which led to remission, and then randomly assigned to two groups. The first group contained 11 patients and received maintenance chemotherapy. The second group contained 12 patients and did not receive maintenance chemotherapy. Time until relapse, the response variable denoted by T hereafter, was recorded in weeks. The objective of the survival analysis was to see if the maintenance chemotherapy prolonged time until relapse. In later sections, we will answer this question by applying appropriate survival analysis methods.

In Group 1, the first and the second patients had relapse times of precisely 9 weeks and 13 weeks, respectively. That is, $T_1 = 9$ and $T_2 = 13$. The third patient, however, was not able to provide an exact

Table 1 The Acute Myelogenous Leukemia Data

Groups	Time to relapse (in weeks)
Group 1 – maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Group 2 – not maintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Note: + = right censoring.

relapse time, perhaps due to being dropped out of the study. All we know is that, at Week 13, this patient was still relapse free. Thus this patient's relapse time should be greater than 13 weeks, which is denoted by 13+ in the table. One can see that for this patient, we have only "incomplete" information because instead of knowing the exact value of T_3 , we know only that $T_3 > 13$. Such censoring is called *right censoring*.

A more complicated censoring mechanism is interval censoring, which arises when the failure time of interest, T , is known only to be bracketed between two observation times, say, $L < T \leq R$. This can happen, for example, in a medical or health study that entails periodic follow-up. An individual who should be observed periodically for a clinically defined "failure" may miss one or more prescheduled visits and return after the failure occurred. In such cases, we know only that the true failure time is greater than the last observation time at which the failure had not occurred and less than or equal to the observation time at which the failure has been observed to occur.

Table 2 presents an interval-censored data set obtained from a cancer study. This study involved 94 early breast cancer patients with two treatments, radiotherapy (radiation therapy) alone and radiotherapy

with adjuvant chemotherapy. Among them, 46 patients were given radiotherapy alone, and 48 patients received radiotherapy with adjuvant chemotherapy. During the study, each patient was supposed to be checked every 4 or 6 months by physicians for the appearance of breast retraction, a response that has a negative impact on overall cosmetic appearance. However, actual observation times differ from patient to patient, and only interval-censored data are available for the appearance. For instance, the second patient in the radiotherapy group had not developed any symptom of breast retraction by Week 25. When she returned at Week 37, however, breast retraction had already appeared. Therefore, the failure time for this patient must be between 25 and 37 weeks. A main objective of the study was to compare the two treatments in terms of the time until the appearance of breast retraction, the survival time of interest.

Interval censoring can be regarded as a more general censoring mechanism than right censoring. If either $L = R - \varepsilon$ for arbitrarily small ε or $R = \infty$, then we have an exact or a right-censored survival time.

Another important feature of many failure time data, sometimes confused with censoring, is *truncation*. Truncation of the failure time occurs when some individuals cannot be observed (and in fact may not even be known to exist) because their failure times do not lie in a certain observational window. This contrasts with censoring, in which case we at least have partial information of the censored observations. Therefore, truncated data must be analyzed under the restriction of conditional inference. In this article, we will focus on survival analysis of censored data and refer readers to some textbooks that discuss truncation in detail.

Kaplan-Meier Estimator and Related Methods

The *Kaplan-Meier (KM) estimation* is a nonparametric technique for estimating the survival function of a homogeneous

Table 2 Breast Cosmetic Deterioration Data

Radiotherapy			Radiotherapy and Chemotherapy		
(45, ∞]	(25, 37]	(37, ∞]	(8, 12]	(0, 5]	(30, 34]
(6, 10]	(46, ∞]	(0, 5]	(0, 22]	(5, 8]	(13, ∞]
(0, 7]	(26, 40]	(18, ∞]	(24, 31]	(12, 20]	(10, 17]
(46, ∞]	(46, ∞]	(24, ∞]	(17, 27]	(11, ∞]	(8, 21]
(46, ∞]	(27, 34]	(36, ∞]	(17, 23]	(33, 40]	(4, 9]
(7, 16]	(36, 44]	(5, 11]	(24, 30]	(31, ∞]	(11, ∞]
(17, ∞]	(46, ∞]	(19, 35]	(16, 24]	(13, 39]	(14, 19]
(7, 14]	(36, 48]	(17, 25]	(13, ∞]	(19, 32]	(4, 8]
(37, 44]	(37, ∞]	(24, ∞]	(11, 13]	(34, ∞]	(34, ∞]
(0, 8]	(40, ∞]	(32, ∞]	(16, 20]	(13, ∞]	(30, 36]
(4, 11]	(17, 25]	(33, ∞]	(18, 25]	(16, 24]	(18, 24]
(15, 8]	(46, ∞]	(19, 26]	(17, 26]	(35, ∞]	(16, 60]
(11, 15]	(11, 18]	(37, ∞]	(32, ∞]	(15, 22]	(35, 39]
(22, ∞]	(38, ∞]	(34, ∞]	(23, ∞]	(11, 17]	(21, 8]
(46, ∞]	(5, 12]	(36, ∞]	(44, 48]	(22, 32]	(11, 20]
(46, ∞]			(14, 17]	(10, 35]	(48, ∞]

Note: ∞ indicates observation is right censored.

right-censored sample. If there is no censoring in the data, we can certainly use the empirical cumulative distribution function (and then subtract it from 1) as our estimator. However, existence of censoring requires more complex treatment.

Let $t_1 < t_2 < \dots < t_k$ represent the exactly observed (i.e., not censored) failure times in a sample of size n . We first estimate the hazard function at time t_i for $i = 1, \dots, k$. Obviously, an intuitive estimator of $\lambda(t_i) = f(t_i)/S(t_i)$ should be $\hat{\lambda}(t_i) = d_i/n_i$, where d_i is the number of failures at time t_i , and n_i is the number of participants at risk just prior to t_i . Given all the $\hat{\lambda}(t_i)$ estimated, the KM estimator for the unknown survival function $S(t)$ is simply $\hat{S}(t) = \prod_{i:t_i \leq t} 1 - \hat{\lambda}(t_i)$. A heuristic argument of the validity of the KM estimator is outlined as follows. Note that we can rewrite $\lambda(x) = f(x)/S(x)$ as $S(t) = \exp[-\int_0^t \lambda(u) du]$. Therefore, a discrete approximation of $S(t)$, if we know only the values of $\hat{\lambda}(t_i)$ at $\{t_1, t_2, \dots, t_k\}$, is $\prod_{i:t_i \leq t} [1 - \hat{\lambda}(t_i)]$.

Confidence intervals of the survival function can be obtained by estimating the variance of $\hat{S}(t)$ via the so-called *Greenwood formula* and assuming asymptotic normality. This topic will not be explored further here.

We now illustrate the above procedure with the AML data in Table 1. For the maintained group, $n = 11$ and $k = 7$. For the other group, $n = 12$ and $k = 9$.

The intermediate quantities we need, together with the final estimate of $S(t)$, are shown in Table 3.

Notice that censored observations contribute to the change of n_i in the above calculation. In the next section, these two (estimated) survival functions will be plotted and contrasted to see whether maintenance prolongs time to relapse.

When interval-censored data are present, the KM estimator can no longer be used because we usually have not any exactly observed failure time (recall that now all failure times fall in some interval). In such cases, an extension of the KM estimator, the *Turnbull estimator*, should be employed. This estimation procedure, which is essentially an expectation maximization (EM)-type algorithm, requires iterative steps and is outlined as follows. Let the observed data be $\{(l_i, r_i]; i = 1, \dots, n\}$ and $0 = a_0 < a_1 < \dots < a_m = \infty$ such that each l_i and r_i is contained in the set. Define $\alpha_{ij} = 1$ if $(a_{j-1}, a_j]$ is a subset of $(l_i, r_i]$ and 0 otherwise for $i = 1, \dots, n$ and $j = 1, \dots, m$. Note that α_{ij} essentially indicates whether participant i contributes to the estimation of the survival function (of T) over the interval $(a_{j-1}, a_j]$, depending on whether it is likely that the (unobserved) failure time of the i th participant falls within this interval. The following Steps 1 to 3 are then iteratively repeated until convergence.

Table 3 Kaplan-Meier Estimates for the AML Data

Group 1 – Maintained						Group 2 – Not Maintained					
i	t_i	d_i	n_i	$\hat{\lambda}(t_i)$	$\hat{S}(t_i)$	i	t_i	d_i	n_i	$\hat{\lambda}$	$\hat{S}(t_i)$
1	9	1	11	1/11	0.909	1	5	2	12	2/12	0.833
2	13	1	10	1/10	0.818	2	8	2	10	2/10	0.667
3	18	1	8	1/8	0.716	3	12	1	8	1/8	0.583
4	23	1	7	1/7	0.614	4	23	1	6	1/6	0.486
5	31	1	5	1/5	0.491	5	27	1	5	1/5	0.389
6	34	1	4	1/4	0.368	6	30	1	4	1/4	0.292
7	48	1	2	1/2	0.184	7	33	1	3	1/3	0.194
						8	43	1	2	1/2	0.097
						9	45	1	1	1/1	0.000

Note: i = index; t_i = i th uncensored observation time point; d_i = number of failures at time t_i ; n_i = number of subjects at risk just prior to t_i ; $\hat{\lambda}(t_i)$ = estimate of the hazard function λ at time t_i ; $\hat{S}(t_i)$ = estimate of the survival function $S(t_i)$ at time t_i .

Table 4 Turnbull Estimates for the Breast Cosmetic Deterioration Data

t	0 – 4	5 – 6	7	8 – 11	12 – 24	25 – 33	34 – 38	40 – 48	≥ 48
$\hat{S}(t)$	1.000	0.954	0.920	0.832	0.761	0.668	0.586	0.467	0.000

Note: t = time to failure (i.e., time to breast cosmetic deterioration); $\hat{S}(t)$ = estimate of the survival function $S(t)$ at time t .

Step 0. Assign initial values to $S(a_j)$ for $j = 1, \dots, m-1$ (recall that $S(a_0) = 1$ and $S(a_m) = 0$). A common choice is $S(a_j) = (m-j)/m$; that is, equal weight is given to each time point.

Step 1. Compute the probability p_j that an event occurs in interval $(a_{j-1}, a_j]$; that is, $p_j = S(a_{j-1}) - S(a_j)$, $j = 1, \dots, m$.

Step 2. Update p_j by

$$p_j^{\text{new}} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ij} p_j}{\sum_{k=1}^m \alpha_{ik} p_k}.$$

Note that the summand

$$\frac{\alpha_{ij} p_j}{\sum_{k=1}^m \alpha_{ik} p_k}$$

represents the probability that the i th failure time is in interval $(a_{j-1}, a_j]$ according to the (old) $p_1, \dots, p_m$.

Step 3. Compare p_j^{new} and p_j for all $j = 1, \dots, m$. If they are close enough, stop. If not, replace p_j by p_j^{new} in Step 1 and repeat Steps 2 to 3.

It can be shown that, under mild conditions, the above algorithm always converges although the convergence can be very slow when sample size is large. Table 4 gives the Turnbull estimate of the survival function for the radiotherapy group of the breast cosmetic deterioration data (see Table 2).

Some programming tips: In S-PLUS, the KM estimate is computed by the routine `Survfit`, and the Turnbull estimate can be obtained from `kaplanMeier`. SAS, however, has only `LIFETEST` for computing the KM estimate.

Log-Rank Tests

Sometimes it is of interest to test whether multiple survival functions are identical. For example, as

described above, a physician may want to know whether receiving maintenance chemotherapy really helps patients. In such a case, the null hypothesis is $H_0: S_1(t) = S_2(t)$ for all t , and the alternative hypothesis is $H_1: S_1(t) \geq S_2(t)$ for all t with at least one strict inequality. We first consider the right-censored AML data and plot the two estimated survival functions obtained earlier.

It seems that the null hypothesis should be rejected according to Figure 1. Nonetheless, we need to develop a test statistic to rigorously justify our conclusion. The *log-rank test*, which may be regarded as an extension of the Mann-Whitney test for uncensored data, is one of the most popular choices.

Let $t_1 < t_2 < \dots < t_m$ be the ordered observed distinct failure times (not censored observations) in the sample formed by combining the two groups, and let D_{1k} and Y_{1k} denote the number of observed failures and number at risk, respectively, for the first group at time t_k , $k = 1, \dots, m$. Also let D_k and Y_k be the corresponding values in the combined sample. Under the

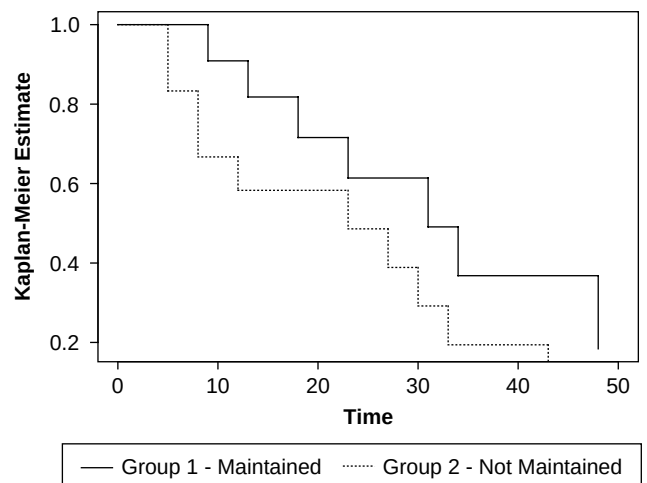


Figure 1 AML Data: Kaplan-Meier Estimate of the Survival Functions

null hypothesis that the two groups share a common survival function (and thus hazard function), given Y_{1k} , D_k and Y_k , D_{1k} has a hypergeometric distribution with mean $E_{1k} = D_k Y_{1k} / Y_k$ and variance $V_{1k} = D_k Y_{1k} (Y_k - Y_{1k}) (Y_k - D_k) / Y_k^2 (Y_k - 1)$. Define the test statistic as

$$Z = \sum_{k=1}^m (D_{1k} - E_{1k}) / \sqrt{\sum_{k=1}^m V_{1k}}.$$

It can be shown that Z approximately follows a standard normal distribution when sample sizes are large. Note that the test statistic Z is essentially a quantity of standardized observed-minus-conditionally-expected number of events.

Applying the above test to the AML data, we obtain $Z = -1.843$, yielding a p value of 0.033 for the one-sided test specified above. Thus we may conclude that the maintenance chemotherapy prolonged time until relapse. If a two-sided test is preferred, that is, the alternative hypothesis is $S_1(t) \neq S_2(t)$ for some t , then the p value would be 0.065.

A few remarks are in order. First, the name *log-rank* arises from two facts. One is that, in computing the test statistic Z , we actually estimated two cumulative hazard functions (one for Group 1 and the other for the combined sample), which are the negative logarithm of the survival functions. The other fact is that what matters is really the ranks of the observed failure times. In other words, if one monotonically transforms the data, say by multiplying all the time points by 10 or taking square roots, the test statistic will remain unchanged.

Second, the log-rank test presented above has many modifications and extensions—so many that sometimes they become confused with each other. One major modification is that one can apply the above method to multiple (rather than only two) samples. Another is that one can specify various weight functions in calculating Z . Furthermore, the log-rank test we have discussed is not applicable if the data are interval censored. For such a case, we refer the readers to the book by J. Sun under “Further Reading” because the test procedures under interval censoring are usually complicated. Several excellent textbooks are also provided as further references.

Last, in S-PLUS, the routine `survdif` can test the hypothesis laid out above. In SAS, the log-rank test is carried out by the `LIFETEST` procedure.

Semiparametric Regression Analysis

When covariates are present, especially when they are ordinal or continuous, regression analysis often needs to be performed. However, nonnegativity of the failure time and censoring complicate the problem. In addition, in medical studies, clinical trials, and many other fields, it is often not appropriate to make parametric assumptions. Consequently, semiparametric regression models are widely used in survival analysis.

The most commonly used semiparametric model, the *proportional hazards (PH) model*, is also termed as the *Cox model* in the literature. Denote the failure time by T as before, and a vector of covariates by $\mathbf{Z}$. The PH model assumes that $\lambda(t | \mathbf{Z} = \mathbf{z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{z}\}$, where $\boldsymbol{\beta}$ is a vector of unknown regression parameters, with $\boldsymbol{\beta}'\mathbf{z}$ denoting the inner product of two vectors $\boldsymbol{\beta}$ and $\mathbf{z}$, and $\lambda_0(t)$ is an unspecified nonnegative hazard function. Here $\lambda(t | \mathbf{Z} = \mathbf{z})$ should be interpreted as the hazard function of T given the covariate $\mathbf{Z} = \mathbf{z}$. Notice that when $\mathbf{Z} = \mathbf{0}$, the hazard function of T is nothing but $\lambda_0(t)$. Therefore, $\lambda_0(t)$ is termed the *baseline hazard function*. Because $\lambda_0(t)$, having an unknown form, is essentially a nonparametric component and $\exp\{\boldsymbol{\beta}'\mathbf{z}\}$ is obviously a parametric one, the PH model is therefore indeed a “semiparametric” model.

Suppose individual i has a covariate vector $\mathbf{Z}_i$ and another individual j has $\mathbf{Z}_j$. Then the two corresponding hazard functions suggest that

$$\begin{aligned} \frac{\lambda_i(t | \mathbf{Z}_i = \mathbf{z}_i)}{\lambda_j(t | \mathbf{Z}_j = \mathbf{z}_j)} &= \frac{\lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{z}_i\}}{\lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{z}_j\}} \\ &= \exp\{\boldsymbol{\beta}'(\mathbf{z}_i - \mathbf{z}_j)\}, \end{aligned}$$

which is a constant. In other words, the two hazard functions are proportional to each other. That is where the model obtains its name.

For the AML data, denote the unmaintained group by $Z = 0$ and the maintained group by $Z = 1$ (here Z is a scalar). The PH model says that time to relapse of those patients without maintained chemotherapy has a hazard

function $\lambda_0(t)$ while that of the patients receiving chemotherapy has $\lambda_0(t)e^\beta$. In other words, the difference caused by the maintained chemotherapy is reflected by e^β (or rather, β), which we will estimate next.

Often we are interested in estimating β , the regression parameter, because it determines how the covariates affect the failure time. An approach called *partial likelihood* enables us to fulfill the task without being troubled by the unknown $\lambda_0(t)$. Suppose there are no ties between the observed failure times (not censored observations). Let $t_1 < t_2 < \dots < t_m$ be the ordered observed failure times, $\mathbf{z}_{(j)}$ be the covariate associated with the individual whose failure time is t_j , and $R(t_j)$ the risk set at time t_j —, that is, the set of all individuals who are still under observation just prior to t_j , $j = 1, \dots, m$. The partial likelihood of the data, based on the PH model, is expressed by the following equation:

$$L(\beta) = \prod_{j=1}^m \frac{\exp\{\beta' \mathbf{z}_{(j)}\}}{\sum_{k \in R(t_j)} \exp\{\beta' \mathbf{z}_k\}}.$$

Maximizing the above partial likelihood or its logarithm, which requires numerical analysis tools, yields the estimator $\hat{\beta}$. Notice that the numerator utilizes the information contributed by the j th failure, whereas the denominator carries the information about all individuals who are still at risk, some of which may be censored later.

The partial likelihood specified above is not the usual full likelihood. It is not even a conditional likelihood, although under mild assumptions it coincides with the marginal likelihood. Nonetheless, it yields an efficient and asymptotic normal estimator of β and can be treated as an ordinary likelihood. In fact, the variance of $\hat{\beta}$ is estimated by calculating the inverted observed information matrix from the partial likelihood.

When ties are present, we need to modify the above partial likelihood to break the tie. Several methods exist in the literature for such a purpose. For example, the Breslow version of the modified partial likelihood is

$$L(\beta) = \prod_{j=1}^m \frac{\exp\{\beta' \mathbf{s}_{(j)}\}}{\left[\sum_{k \in R(t_j)} \exp\{\beta' \mathbf{z}_k\} \right]^{d_j}},$$

where d_j is the number of failures at time t_j and $\mathbf{s}_{(j)} = \sum_{i: t_i = t_j} \mathbf{z}_i$ is the sum of the covariates of individuals observed to fail at t_j .

Applying the PH model to the AML data by coding Z as mentioned earlier, we obtain $\hat{\beta} = -0.904$ with a standard error of 0.512. Thus the p value for testing $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ is 0.078 (note that this result is consistent with the log-rank test result), indicating marginal significance. The value of $\hat{\beta}$ can also be interpreted via the hazard ratio—we may say that the hazard function corresponding to the maintained group is roughly 40% ($e^{-0.904} = 0.405$) of the hazard function corresponding to the unmaintained group. Put another way, if two patients, one in the maintained group and the other in the unmaintained group, both have survived up to time t , then the probability that the former will “fail” (in this case, relapse) in the next moment is only about 40% of the probability that the latter will.

It must be emphasized that one should ignore the baseline hazard function $\lambda_0(t)$ only when the interest of study lies solely in testing whether the covariate influences the failure time. If one also wants to predict the survival probability for a particular patient, then $\lambda_0(t)$ will need to be estimated too. The most commonly used estimator is the Aalen-Breslow estimator.

When interval censoring is present, things are much more complicated. The reason is that we will in general have to estimate the baseline hazard function $\lambda_0(t)$ simultaneously when we estimate β because the partial likelihood approach no longer applies. One remedy is to consider a more general type of semiparametric model, termed linear transformation models, which contain the PH model as a special case.

The SAS procedure PHREG contains many extensions of the PH model. The corresponding routine in S-PLUS is `coxph`.

—Zhigang Zhang

Further Reading

- Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89–99.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society B*, 33, 187–220.

- Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. Hoboken, NJ: Wiley.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.
- Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of Royal Statistical Society A*, 135, 185–206.
- Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.
- Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of Royal Statistical Society B*, 38, 290–295.
- Zhang, Z., Sun, L., Zhao, X., & Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33, 61–70.

SYSTEM OF MULTICULTURAL PLURALISTIC ASSESSMENT

The System of Multicultural Pluralistic Assessment, best known as SOMPA, was developed by S. Mercer and A. Lewis in 1978 to meet the educational and mental health needs of minority children. Published by the Psychological Corporation, the SOMPA was discontinued in 2003. The SOMPA was designed as a system for assessing the level at which children function in cognitive abilities, perceptual motor abilities, and adaptive behavior in a nondiscriminatory manner.

Intended to provide a comprehensive assessment of children ages 5 to 11, the SOMPA included a medical component, a social component, and a pluralistic component. The medical component determined the presence or absence of organic pathology. Six measures were used to assess the medical component: physical dexterity tasks (sensory motor coordination), the Bender Visual Motor Gestalt Test (perceptual and neurological factors), the Health History Inventories, weight by height norms (nutritional or developmental problems), vision (the Snellen Test), and auditory acuity (national norms).

The social component was concerned with a child's ability to adapt and function in social situations. Two measures were used to assess the social component: the Adaptive Behavior Inventory for Children and the Wechsler Intelligence Scale for Children-Revised (WISC-R).

Last, the pluralistic perspective yielded an index of a child's intelligence or Estimated Learning Potential through a "corrected" WISC-R score based on a comparison of how well the child performed on that test with the performance of other children who had had similar learning opportunities.

The procedures of the SOMPA provided an alternative testing method to increase the proportion of minority students in gifted education programs, particularly in states that used IQ cut-off scores for placement decisions. There was some evidence to suggest that IQ or Estimated Learning Potential, as measured by the SOMPA, was able to predict school achievement. In addition, the SOMPA had some utility as an aide to psychological assessment of Native American Navajo children. In spite of the general acceptance of the SOMPA as a tool for assessing children from culturally and linguistically diverse backgrounds, some criticism remained unresolved. One was the claim that the SOMPA was a better instrument than the WISC for predicting learning potential in a school environment. Also, concerns were raised regarding the national representativeness of the sample. Despite the criticism, the SOMPA represented the first attempt at reducing cultural biases and stigmatization in assessing culturally and linguistically diverse children.

—Romilia Domínguez de Ramírez

Further Reading

- Figuerroa, R. (1982). SOMPA and the psychological testing of Hispanic children. *Metas*, 2, 1–6.
- Figuerroa, R. A., & Sassenrath, J. M. (1989). A longitudinal study of the predictive validity of the System of Multicultural Pluralistic Assessment (SOMPA) (ERIC Document Reproduction Service No. EJ391800). *Psychology in the Schools*, 26, 5–19.

T

The only man I know who behaves sensibly is my tailor; he takes my measurements anew each time he sees me. The rest go on with their old measurements and expect me to fit them.

—George Bernard Shaw

***T* SCORES**

A *T* score is a type of standard (not standardized) score that has a mean of 50 and a standard deviation of 10. It is very similar in concept to a *z* score, which has a mean of 0 and a standard deviation of 1.

T scores are used when the researcher wants to convert raw scores to a metric that is comparable across distributions (different sets of scores) and where there is a desire to have all scores as positive values (which is the case unless a raw score is more than 3 standard deviations below the mean).

The formula for *T* score is

$$T = 50 + 10z,$$

where

T is the *T* scores,

50 is the mean of the set of *T* scores,

10 is the amount that one *T* score deviates from the mean, and

z is the corresponding *z* score for a particular raw score.

For example, here is a set of raw scores and their corresponding *z* scores and *T* scores:

Raw Score	<i>z</i> Score	<i>T</i> Score
6	1.17	61.65
4	−0.78	42.23
5	0.19	51.94
4	−0.78	42.23
3	−1.75	32.52
6	1.17	61.65
5	0.19	51.94
4	−0.78	42.23
5	0.19	51.94
6	1.17	61.65

You can see that a raw score with a *z* score close to the mean of the distribution (which is 4.8) is the closest to 50 and that all the *T* scores are well within the positive range.

—Neil J. Salkind

See also Standard Scores; *z* Scores

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

***t* TEST FOR TWO POPULATION MEANS**

The *t* test for independent means examines the difference between the means of two independent groups and requires that for each case in the sample, there be two variables. The first is the *group variable* (such as treatment or gender), and the second is the *test variable* (such as a score on a personality test or an achievement test). The second variable, sometimes known as the *grouping variable*, places each individual in one of two mutually exclusive categories, and the *t* test itself evaluates whether there is a significant difference between the two groups.

Why is it referred to as “Student’s *t*?” The test was formulated by William Gossett (a student of Karl Pearson) in the early 1900s, when he was a chemist and a statistician at the Guinness Brewing Company. As is true today, many company secrets were proprietary, and his employer would not allow him to publish his own work under his own name (trade secrets and so forth). Instead, he was given permission to publish it under the pseudonym “Student.”

The Case Study and the Data

It is not difficult to find any school system that relies on testing for a variety of different purposes, such as adhering to different federal and state guidelines, for example, the No Child Left Behind Act of 2001. Another purpose might be to chart school progress or determine differences between classrooms at the same grade level. To examine whether such differences are significant, a *t* test for independent means can be applied. Table 1 shows the sample data set for 25 children, one set from Susan Graves’s classroom and one set from Jack Longer’s classroom.

The Assumptions Underlying the *t* Test

Three important assumptions underlie the use of the *t* test for independent means:

Table 1 Sample *t* Test Data

<i>Student</i>	<i>Graves</i>	<i>Longer</i>
1	89	47
2	89	67
3	85	78
4	76	89
5	68	87
6	95	65
7	99	67
8	87	62
9	67	51
10	76	69
11	92	56
12	89	99
13	85	97
14	65	80
15	72	86
16	78	70
17	76	78
18	77	72
19	53	67
20	78	99
21	91	86
22	67	54
23	80	
24	87	
25	66	

1. The test variable (which in this example is the math test score for each student) is normally distributed. If the sample is large enough (15 or more per group), this assumption is fairly resistant to being violated, but if the scores are not normal, then a larger sample size might be needed.
2. The variances for each of the test variables in both groups are equal to one another.
3. The cases in each of the samples are random in nature, and the scores on the test variable are independent of one another. If this assumption is violated, then the resulting *t* value should not be trusted.

The Research Hypothesis

The null hypothesis associated with this analysis is that there is a difference between the population

means of the two samples. This is a nondirectional test and can be stated as follows:

$$H_1 : \mu_1 = \mu_2,$$

where μ equals the population mean.

Computing the t Value

One formula for computing the t value is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left[\frac{n_1 + n_2}{n_1 n_2} \right]}},$$

where

$\bar{X}_1$ is the mean for Group 1, which in this example is 79.48;

$\bar{X}_2$ is the mean for Group 2, which in this example is 73.91;

n_1 is the number of participants in Group 1, which in this example is 25;

n_2 is the number of participants in Group 2, which in this example is 22;

s_1^2 is the variance for Group 1, which in this example is 11.17;

s_2^2 is the variance for Group 2, which in this example is 15.32.

Once the appropriate values are entered into the equation, it appears like this:

$$t = \frac{79.48 - 73.91}{\sqrt{\left[\frac{(25 - 1)11.17^2 + (22 - 1)15.32^2}{25 + 22 - 2} \right] \left[\frac{25 + 22}{25 \cdot 22} \right]}} = 1.436.$$

Using the Computer

For this example, SPSS was used to compute the t value for this test of independent means. The output shown in Figure 1 includes the mean standard deviation and standard error of the mean for the dependent variable (math score) for both the Graves and Longer classrooms. Also shown in the output is the computation of the t value (equal to 1.436), the degrees of freedom (45), and the associated and exact level of significance (.158). Other information in this output may be useful as well.

Group Statistics

	Class	N	Mean	Std. Deviation	Std. Error Mean
Score	Graves	25	79.4800	11.17333	2.23467
	Longer	22	73.9091	15.31855	3.26593

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Score	Equal variances assumed	2.964	.092	1.436	45	.158	5.57091	3.87914	-2.24207	13.38389
	Equal variances not assumed			1.408	37.982	.167	5.57091	3.95727	-2.44030	13.58211

Figure 1 Group Statistics, Independent Samples Test

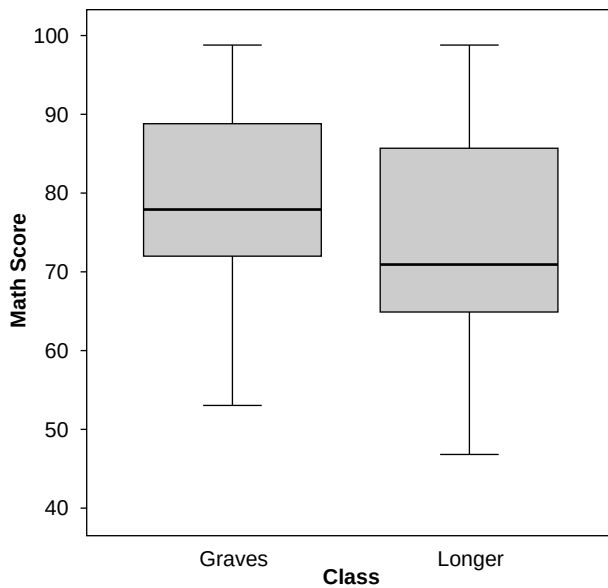


Figure 2 Means and Standard Deviations, Range of Scores

The chart shown in Figure 2 is a box plot chart created using SPSS and shows the means and standard deviations for each of the two groups as well as the range of scores.

—Neil J. Salkind

See also Ability Tests

Further Reading

- Bean, P., Loomis, C., Timmel, P., Hallinan, P., Moore, S., Mammel, J., et al. (2004). Outcome variables for anorexic males and females one year after discharge from residential treatment. *Journal of Addictive Diseases*, 23, 83–94.
- Clark, R. (2004). Interethnic group and intra-ethnic group racism: Perceptions and coping in Black university students. *Journal of Black Psychology*, 30, 506–526.

Student's *t* Test Applet: <http://nimitz.mcs.kent.edu/~blewis/stat/tTest.html> (allows you to enter data and solve for *t*)

Applying Ideas in Statistics and Measurement

The following abstract is adapted from Heydenberk, R. A., & Heydenberk, W. R. (2005). Increasing meta-cognitive competence through conflict resolution. Applying ideas on statistics and measurement. *Education and Urban Society*, 37(4), 431–452.

The ***t* test** is one of the most popular of all inferential tests, since the comparison it does between two groups (be they independent or dependent) is such a common comparison found in almost any science. In this study, Heydenberk and Heydenberk, from Lehigh University, determined the effects of conflict resolution and related social skill development on students' metacognitive competencies. The study was conducted over a 5-year period in elementary schools in the Philadelphia School District and in a neighboring urban school district, and the participants were fourth and fifth graders. The performance of each student was assessed for significant pretest-to-posttest differences using a one-tailed ***t* test** with an alpha or significance level of .05. The results showed that students who received the treatment demonstrated significant improvement in metacognitive skills, and the research hypothesis was accepted—that the introduction of conflict resolution worked. Their conclusion? Integrating conflict resolution and social skills training into curricula is an effective way to affect students' metacognitive competences.

TEST-RETEST RELIABILITY

All observations and all measurements contain error. The focus of much work in measurement is on minimizing and estimating the amount of error in any given measurement. In classical test theory, *X* is an observed score that is composed of *T*, the true score, and *E*, the error score: $X = T + E$. The true score is never known, but it can be thought of as the long-range average of scores from a single instrument administered to an individual an infinite number of times (the expected value or expected score). The error score is random and may have many sources, including testing conditions, individual characteristics that fluctuate from administration to administration, differences in forms, and instability of an individual's ability or trait over time.

This random error score is quite different from systematic sources of error, like testwiseness, which may systematically increase an individual's score on each administration. Since testwiseness is systematic or

constant, it finds its way into the true score and creates problems regarding validity, since the trait being measured may be inadvertently influenced by testwiseness. Random error, since it varies randomly, influences the consistency of scores but not the expected value of a score (the true score) and thus influences reliability, not validity.

Theoretically, we can estimate the amount of error if we know how much of a given score is due to errors of measurement. If we were able to test a single person repeatedly without the effects of recall and fatigue, variation in that person's scores would be considered measurement error. If there were no measurement error, that person would get the same score on each administration. Since it is not possible to test individuals repeatedly without the interference of recall and fatigue, we employ groups to estimate measurement error variance. This allows us to estimate the *standard error of measurement*, the typical amount of measurement error in a set of scores.

If we take the classical test theory model of scores and consider groups of scores and their variances, we see that the variance of the observed scores equals the sum of the variance of true scores and the variance of error scores: $S_X^2 = S_T^2 + S_E^2$ (in sample notation).

This is the long way of introducing the need for reliability. *Reliability* is a tool used to estimate the standard error of measurement but also has some intrinsic benefits in and of itself. Theoretically, reliability is considered the correlation between scores on two parallel forms of a test. The idea is that if there is no measurement error at work, scores from two parallel forms administered to the same group of individuals should be perfectly correlated—each individual should obtain the same score. It can be shown that the correlation between two parallel forms of a test is equal to the ratio of true score variance to observed score variance—the proportion of variance in observed scores that is due to true individual differences: $r_{tt} = \frac{S_T^2}{S_X^2}$. This reliability coefficient can then be used in estimation of the standard error of measurement, because it tells us the proportion of observed variance that is *true variance*; the standard error of measurement is a function of the proportion of observed variance that is true variance.

Estimating Reliability

Based on the classical test theory conception of reliability, scores are reliable to the extent that individuals' scores remain constant on repeated measurements. One way to estimate the degree of consistency or stability of scores is to administer a test on multiple occasions. Each administration provides a separate observation or measurement, from which stability of scores can be observed directly.

This form of reliability is best estimated when the actual measurement procedure employs multiple administrations of a single form, because if scores vary between administrations for a single individual, we are likely to attribute this to differences in occasions or time sampling—errors result from conditions in the lapse of time. This is typically done by administering the test form to each individual twice, within a specified time period. The scores resulting from the two administrations are then correlated—this correlation is the *test-retest reliability coefficient*.

Formal Conditions for Parallel Forms Reliability

Test-retest reliability is estimated from the administration of one form on two occasions. This is appropriate when the measurement procedure includes a single form that may be administered on multiple occasions (pre-post tests or follow-up administrations) or the inference from scores is based on the ability to generalize in terms of the occasion or time frame of measurement. Knowledge of test-retest reliability allows one to generalize from one time of measurement to other time periods or occasions. For this to be a coefficient of reliability, the test should be administered in a specified time frame and occasion to the degree that is meaningful given the characteristics being measured. To be able to adequately interpret a test-retest reliability coefficient, one must know the length of time between administrations, the differences in specific occasions (time of day, day of week, etc.), and the existence of relevant intervening experiences that are likely to result in changes in scores.

At least three sources of score variability will lead to instability of scores over time. First, variation due to instability of the trait or ability being measured from time to time will lead to variation in scores. Variation in attention, motivation, and so on will lead to variation in scores. Finally, if the forms differ from one administration to another, content sampling will also introduce variation in scores (often referred to as a *stability-equivalence estimate* of reliability). To the extent that the trait is normally thought of as being unstable or easily malleable by intervening experiences, test-retest reliability would not be an appropriate method of estimating reliability. If we expected an intervention to alter scores and needed to estimate measurement error, an estimate of test-retest reliability would overestimate measurement error, since we produced a condition that created changes in scores (the intervention). Test-retest reliability is an appropriate estimate of reliability when the trait is generally expected to be stable over time and we are interested in estimating the stability of scores from an instrument intended to measure that trait.

Test-retest reliability coefficients are typically lower when the administrations occur over longer periods of time, when there are more opportunities for relevant intervening experiences that change the trait being measured, and when the content of the forms changes. At the same time, if a given form of a test is administered within a relatively short period of time, the effects of memory or recall are likely to influence scores in a way that is difficult to estimate. Recall will prevent the individual from responding in a truthful way and biasing results.

General Issues Regarding Reliability

Because the test-retest reliability estimate is based on a correlation, it is not simply a characteristic of the measurement instrument itself. Score variability directly influences correlations, such that all else being equal, the more score variance present, the higher the correlation and thus the higher the reliability. Correlational forms of reliability are sample specific and thus are not necessarily generalizable to other samples. They do, however, provide an estimate of score consistency for the scores at hand.

In any estimate of reliability, conditions present during the specific administration of the measurement instrument can impact performance and scores in random ways, leading to lower consistency of scores and lower reliability. Each type of reliability estimate (e.g., test-retest reliability) also captures a specific form of random error. The test-retest reliability primarily captures measurement error due to sampling time or occasions. If this source of error is important to estimate given the measurement procedure (because the test is administered multiple occasions over time), then it is an appropriate form of reliability. Technically speaking, an estimate of reliability should be obtained for each set of scores, since any one estimate is sample specific and the argument of generalizability across samples is difficult to make.

Finally, because sampling error is a function of sample size, all else being equal, longer forms will yield higher reliability coefficients. Better, larger samples of items from the domain will reduce the likelihood that two forms differ in their ability to cover the domain. A functional relation between form length and reliability is represented by the Spearman-Brown prophecy formula.

—Michael C. Rodriguez

See also Coefficient Alpha; Reliability Theory; Standard Error of Measurement

Further Reading

- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education; Macmillan.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6), 956–972.

Of the several different types of reliability, test-retest is very often used to establish the consistency

of a measure. A total of 1,247 college students participated in this study on the effect of scale format on the reliability of Likert-type rating scales. The number of response categories ranged from 3 to 9. The results indicated that the scales with few response categories tended to result in lower reliability, especially lower **test-retest reliability**. The scales with all the response options clearly labeled were likely to yield higher test-retest reliability than those with only the end points labeled. Scale design that leads to consistent participant responses as indicated by test-retest reliability should be preferred.

TESTS OF MEDIATING EFFECTS

Mediator variables are common in disciplines such as psychology, sociology, management, education, political science, and public administration. A mediator variable (referred to hereafter as a *mediator*) transmits the effects of an independent variable to a dependent variable. This is illustrated in Figure 1a, which shows a causal chain involving an independent variable (Z_1), a mediator (Z_2), and a dependent variable (Z_3). In instances of actual (as opposed to assumed) mediation, two types of effects are possible. We illustrate these with reference to Figures 1a and 1b. In *complete mediation*, there is only an indirect effect of the independent variable on the dependent variable (see Figure 1a). However, in *partial mediation*, an independent variable has both a direct effect on the dependent variable and an indirect effect (see Figure 1b).

There are many examples of mediating effects. For instance, behavioral intentions are assumed to mediate the relation between attitudes and behavior, and stress is hypothesized to mediate the relation between stressors and strain.

Research Design Issues and Tests of Mediation

Tests of mediation may be based on data from studies using three major types of designs: randomized experiments, quasi-experiments, and nonexperiments. Inferences about mediation that stem from such tests

rest on a relatively firm foundation when they are based on data from randomized experiments. For example, in testing for the mediating effect of Z_2 on the relation between Z_1 and Z_3 , one randomized experiment can be used to show that Z_1 causes Z_2 , and another can be conducted to demonstrate that Z_2 causes Z_3 .

Unfortunately, when tests of mediation are based on data from quasi-experimental studies, inferences about mediation rest on a much weaker foundation. Moreover, when such tests rely on data from non-experimental research, inferences about mediating effects are almost never justified. One of the major

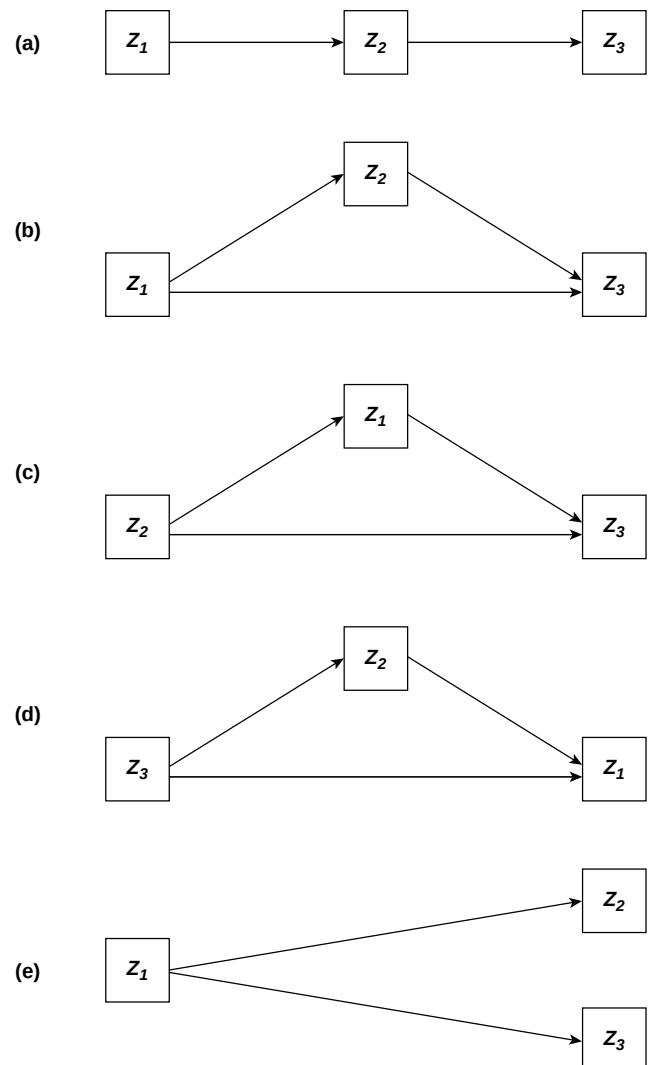


Figure 1 Some Possible Causal Relations Among Three Variables

Source: Stone-Romero & Rosopa (2004).

reasons for this is that when tests of mediation are based on data from nonexperimental research, all that a researcher can confidently conclude is that an observed pattern of relations (e.g., covariances, correlation coefficients) among the assumed independent, mediator, and dependent variables is consistent with a hypothesized model (e.g., that shown in Figure 1b). However, the same pattern of relations may also be consistent with a number of alternative models, including those shown in Figures 1c to 1e. Thus, the results of nonexperimental research almost always provide highly equivocal evidence about mediation. Moreover, as noted below, the validity of inferences about mediation is not at all improved by the use of various statistical procedures that purport to test causal models. One important reason for this is that in quasi-experimental or nonexperimental research, mediation models *assume* a specific set of causal connections between (among) variables (e.g., the model shown in Figure 1b), and statistical procedures are incapable of providing credible evidence of the correctness of the assumed pattern of relations.

Statistical Approaches for Testing Mediation Models

A number of statistical procedures have been used in tests of mediation that are based on data from nonexperimental or quasi-experimental research. Among these are *path analysis* (PA), *structural equation modeling* (SEM), and *hierarchical multiple regression* (HMR). Contrary to seemingly very popular beliefs, none of these provide valid tests of mediation. Reasons for this are detailed below.

Failure to Satisfy Requirements for Causal Inference

Three major conditions are vital to demonstrating a causal connection between a hypothesized cause and an assumed effect: (a) The hypothesized cause must precede the presumed effect in time; (b) the variables must be related to one another; and (c) the relation between the variables must not be a function of one or more alternative causes. These conditions are easy to satisfy in well-designed and well-conducted

experimental studies. However, in quasi-experimental research, only the first two conditions can be satisfied. And in nonexperimental research, only the second condition can be met. As a result, when techniques such as PA, SEM, and HMR are applied to data from quasi-experimental and nonexperimental research, there is no valid basis for inferences about causal connections among the hypothesized independent, mediator, and dependent variables. It is critical to recognize that just because variables are considered to be independent, mediator, and dependent in statistical analyses does not mean that they actually have this status. Regrettably, many researchers in psychology and other social sciences seem to have little understanding of or appreciation of this. As a result, they very frequently make unwarranted and inappropriate inferences about mediating effects using data from either quasi-experimental or nonexperimental studies.

It deserves adding that even individuals who advocate using statistical techniques to detect mediating effects recognize that inferences about mediation are problematic when they are based on data from nonexperimental research. For example, in 2002, MacKinnon et al. observed that the HMR procedure does not provide “the full set of necessary conditions for strong inference of a causal effect of the independent variable on the dependent variable through the intervening [mediator] variable.” However, they argued that it and similar procedures provide suggestive evidence on mediation. In addition, in 2002, Shrout and Bolger noted that “statistical analyses of association cannot establish causal links definitively.” Despite this, they stated that such analyses can “provide evidence that one mediation pattern is more plausible than another.” For the reasons noted above, Stone-Romero and Rosopa disagree strongly with this view.

Model Specification Issues

One of the major reasons that statistical tests of mediating effects are incapable of providing valid inferences about mediation is that *model misspecification* is always an issue in such tests. Several aspects of the misspecification problem are discussed below.

Alternative Causal Models

Tests of mediation models that are based on data from quasi-experimental or nonexperimental research assume a particular causal chain (e.g., Figure 1b). However, the actual causal sequence may be different from that which is specified by a researcher. For example, instead of the actual causal chain being the one illustrated in Figure 1b, it may be one of the other models shown in Figure 1 (e.g., the one in Figure 1c). This is an extremely important issue, because when HMR is applied to a set of covariances among three variables (e.g., Z_1 , Z_2 , and Z_3) that stem from quasi-experimental or nonexperimental research, the results of the analysis can often provide support for an incorrect causal model. The same is true of other statistical techniques (e.g., PA, SEM) that are frequently used to test assumed causal models. The fact that these techniques are used for the purpose of what is called *causal modeling* does not imply that the results of model testing provide a valid basis for inferring the validity of a hypothesized set of causal connections between (among) variables.

Model misspecification may manifest itself in several ways. We illustrate this with respect to the HMR technique (described below). In the interest of brevity, we comment on only four model misspecification problems here. First, variables in a model may be measured with error, and techniques such as HMR are incapable of addressing this issue. As a result, estimates of “effects” may be biased. Second, models tested by HMR assume that the causal flow is unidirectional. They have no capacity to model reciprocal causation. Third, HMR is typically used to test mediation models that involve a single mediator. However, there may be multiple mediators of relations between independent and dependent variables. Fourth, and finally, HMR-based tests of mediation do not control for the effects of confounding variables.

A Specific Strategy Used for Testing Mediation

One of the most commonly used techniques for testing hypothesized models involving mediation is the Baron and Kenny HMR procedure. It involves three

sequential multiple regression analyses, which are illustrated here, with reference to the model shown in Figure 1b.

Prior to describing the three analyses associated with HMR tests of mediation, it is important to note that the squared multiple correlation coefficients (R^2 s) for each must be nonzero. Unless they are, the results of the analyses cannot be used to infer the existence of relations between variables in models that are tested.

The Hierarchical Multiple Regression Analytic Strategy

In the first analysis, the hypothesized mediator (Z_2) is regressed on the assumed independent variable (Z_1). This analysis must show a statistically significant (i.e., nonzero) regression coefficient for Z_1 ($\beta_{2,1}$), supposedly implying that Z_1 causes Z_2 . In the second analysis, Z_3 is regressed on Z_1 . The analysis must produce a nonzero regression coefficient for Z_1 ($\beta_{3,1}$), ostensibly implying that Z_1 causes Z_3 . In the third analysis, Z_3 is regressed simultaneously on Z_1 and Z_2 , producing regression coefficients for both predictors. This analysis must show that the regression coefficient for Z_1 is lower than the coefficient for the same variable that resulted from the first analysis. Baron and Kenny argue that this provides a basis for inferring the existence of either partial or complete mediation. An inference of *partial mediation* results when the regression coefficient for Z_1 is nonzero in the third analysis, supposedly meaning that some of the effect of Z_1 on Z_3 is direct (nonmediated). The existence of complete mediation is inferred if the regression coefficient for Z_1 does not differ significantly from zero in the third analysis.

Problems With the Hierarchical Multiple Regression Technique

To illustrate the problems associated with the HMR technique, Stone-Romero and Rosopa conducted a simulation study in which four variables were manipulated: (a) The values of correlations among independent, mediator, and dependent variables in the model tested were set at values between .10 and .90; and (b) sample size (N) took on one of 13 values that ranged from 20 to 3,000. The manipulations resulted in over 8,400 unique combinations of correlation coefficients

and sample sizes. For each such combination, the HMR technique was used to test for mediation. The Type I error rate for these analyses was .05.

Note that two scenarios are possible for each test of mediation. One is that the model being tested (i.e., the one shown in Figure 1b) was properly specified; that is, it represented the true causal model underlying the set of correlation coefficients. The other scenario is that the model tested was misspecified; that is, the true causal model was the one shown in Figure 1e, but the model tested was the one depicted in Figure 1b.

Assuming that the model tested was *properly specified*, the HMR procedure provided evidence suggestive of (a) partial mediation in 3,647 cases (43.09% of all HMR analyses) and (b) complete mediation in only 630 cases (7.44% of all HMR analyses). Thus, even if one assumed that (a) the model being tested was properly specified and (b) data from nonexperimental research could serve as a legitimate basis for making causal inferences, the HMR procedure proved to be of relatively low value as a basis for inferring partial mediation. In addition, it provided support for inferences of complete mediation at a rate that was just slightly above the Type I error rate. It is critical to recognize, however, that if the model tested were misspecified, all inferences about mediation would be invalid. In view of the results of their simulation and issues associated with causal inference, Stone-Romero and Rosopa concluded that the HMR technique should not be used to test for mediation.

Other Research on the Power of the HMR Technique

Stone-Romero and Rosopa are not alone in demonstrating the low power of the HMR technique. For example, a 2002 simulation study by MacKinnon et al. showed that it had very low power when mediator effect sizes were small or when medium and sample size were under 100. Thus, they recommended that the difference in regression coefficients strategy be used to test for mediation, rather than the HMR technique. In addition, a 2002 simulation study by Shrout and Bolger showed that the usual HMR-based test of indirect effects had very low power. As a result, they

recommended that bootstrap procedures be used to develop a confidence interval around the estimate of the supposed indirect effect. Their simulation showed that this interval was superior to the interval based on ordinary least squares regression in terms of providing evidence of indirect effects.

Appropriate Strategies for Testing Causal Models Involving Mediation

Clearly, research shows that statistical procedures other than the HMR strategy have improved statistical power in supposed tests of mediation. Nevertheless, we believe that the most important objective in research aimed at showing mediating effects is demonstrating that the causal model hypothesized by a researcher (e.g., Figure 1a or 1b) is more plausible than alternative causal models (e.g., Figures 1c, 1d, and 1e). Statistical techniques have no potential whatsoever to accomplish this goal. Thus, it is recommended that the validity of causal models involving mediation be demonstrated through well-designed and well-executed experimental studies.

—Eugene F. Stone-Romero and Patrick J. Rosopa

See also Moderator Variable; Variable

Further Reading

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18, 115–126.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Rogelberg, S. (Ed.). (2007). *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Stone-Romero, E. F. (2002). The relative validity and usefulness of various empirical research designs. In S. G. Rogelberg

(Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 77–98). Malden, MA: Blackwell.

Stone-Romero, E. F., & Rosopa, P. (2004). Inference problems with hierarchical multiple regression-based tests of mediating effects. *Research in Personnel and Human Resources Management*, 23, 249–290.

TEXT ANALYSIS

Text analysis constitutes a variety of social science research methods designed to ascertain meaning and bring structure to large amounts of unstructured information derived from different acts of communication embodied in written language. As such, text analysis differs from observation-based methods in that the acts in question have already taken place. Examples of texts include books, transcripts of political speeches, advertisements, interviews, newspaper editorials, and government documents. *Text analysis* can be generally defined as a research technique designed to make systematic and replicable inferences from texts. Text analysis is regularly utilized in a number of social science disciplines, including sociology, anthropology, psychology, and political science. With its focus on written language, text analysis is somewhat distinct from *content analysis*, which can focus on text but also encompasses other diverse symbolic behavior such as music, visual art, and body language.

Jack Sprat could eat no fat; his wife could
eat no lean.

And so betwixt them both, they licked the
platter clean.

The meanings inferred from written sources via text analysis may be either manifest or latent. *Manifest meaning* refers to the literal, surface meaning of the words in question. It is a recognition of the universal meaning of the words and phrases employed in the text, divorced from the intentions of the author and/or any reception from its intended audience. With this meaning in mind, one may infer from the age-old nursery rhyme above an equitable

division of a meal by a married couple. By contrast, it is the very situational context—the creator’s intent, the recipient’s response, the connection between the parties—that constitutes the latent meaning of a text. In this case, meaning can be inferred from factors such as the type of language employed (e.g., formal versus informal) or the identity and relative status of the author and of the addressees (e.g., government and citizens). That “Jack” and “his wife” refer to King Charles I and his queen and “the platter” refers to England’s treasury denotes a much more pointed outcome than the amicable repast the surface meaning implies.

Text analysis may be qualitative or quantitative in nature, depending on the orientation of the researcher and the research questions devised. Qualitative analyses tend to be more inductive and exploratory, whereby researchers examine data looking for patterns, while quantitative analyses are more deductive and used to confirm existing hypotheses. Often, both quantitative and qualitative methods are employed within a single research design.

A relatively recent innovation has been the introduction of computer software programs (e.g., ATLAS/ti, NUD*IST) designed to facilitate text analyses by cutting down or even eliminating many of the time-consuming elements, such as the coding of data by hand. In addition, the amount of readily available textual information produced and disseminated digitally by newspapers, academic journals, government publications, businesses, and so on continues to grow exponentially, furnishing a continuous and easily accessible supply of raw data for analysis.

—Johnny Holloway

See also Content Validity

Further Reading

- Popping, R. (2000). *Computer assisted text analysis*. London: Sage.
- Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Erlbaum.

THEMATIC APPERCEPTION TEST

The Thematic Apperception Test (TAT) (author Henry A. Murray; early coauthor Christiana Morgan) was published by Harvard University Press (<http://www.hup.harvard.edu>) in 1943 and is currently distributed by Harcourt Assessment, Inc. (<http://www.harcourtassessment.com>). It is a widely used performance-based personality assessment technique consisting of 31 achromatic cards. The cards contain scenes portraying either a solitary individual, individuals in diverse interpersonal situations, or landscapes. The pictures vary in their levels of ambiguity and detail. For example, one card features a young woman standing in the foreground of a rural scene with two individuals in various poses behind her, and another card is entirely blank. Examiners typically choose a subset of the cards to present, and the instructions may vary. Generally, however, the examinee is asked to create a story about each scene that includes a description of what is happening in the picture, what led up to this event, what the characters are thinking and feeling, and what the outcome is. The stories are often considered to contain projected material reflecting an individual's drives, motives, conflicts, needs, emotions, and other personality dynamics. The test generally takes about 1 to 2 hours to administer, depending on the number of cards that are chosen for presentation, and can be completed in one sitting or across time.

Because there are a number of ways the test is administered and scored, examiners may use a variety of methods for interpreting the stories produced by examinees. They may rely on clinical inference, standardized ratings systems, or both. A number of quantitative rating methods exist for use in clinical practice or in research, and although there is little adequate reliability and validity data available for the TAT itself, many of the scoring systems report their own psychometric properties. Examples of existing scoring systems include those that measure an individual's object relations, ego defense mechanisms, communication deviance, problem solving, and motives.

Although the TAT was designed for use with both children and adults, additional versions of the test

(i.e., the Children's Apperception Test and the Senior Apperception Test) have been created for more specific populations. Others have designed TAT cards for use with specific racial groups to address a concern about the TAT's cross-cultural applicability, though these versions do not appear to be widely used.

—A. Jill Clemence

See also Personality Assessment Inventory; Personality Research Form; Personality Tests

Further Reading

- Gieser, L., & Stein, M. I. (1999). *Evocative images: The Thematic Apperception Test and the art of projection*. Washington, DC: American Psychological Association.
- Murray, H. A. (1943). *Thematic Apperception Test: Manual*. Cambridge, MA: Harvard University Press.
- Robinson, F. (1992). *Love's story told: A life of Henry A. Murray*. Cambridge, MA: Harvard University Press.

Henry A. Murray and Christiana D. Morgan biographies: <http://www.mhhe.com/mayfieldpub/psychtesting/profiles/>

THREE-CARD METHOD

The three-card method is a technique for gathering information from people on questions that are sensitive or controversial in nature. Some topics of interest to researchers and policymakers concern illegal activities, unpopular opinions, and embarrassing personal details, such as alcohol and drug abuse, criminal activity, and sexual behavior. Respondents, if asked directly about sensitive topics, might refuse to answer or might provide false answers.

Instead of being asked a sensitive question directly, a respondent in the three-card method is shown one of three cards. Each card lists valid response categories organized into three distinct groups (A, B, or C). If a response category is sensitive, it is never the only category in a group on a card. That is, the sensitive response is matched with one or more nonsensitive responses on each card. Once shown a card, the respondent tells the interviewer the group to which he or she belongs. The

privacy of the respondent is protected, because the respondents never have to definitely identify themselves as belonging to the controversial group. For example, if foreign individuals in the United States are being asked about immigration status, one possible card would list both “green-card status” (a legal residency status) and “undocumented immigrant” (not a legal resident) in Group B. A second card could list “temporary work visa” (another legal residency status) with “undocumented immigrant” together as Group B. A respondent picking Group B on either card has not directly admitted to being in the United States illegally.

In the whole survey, three cards (Card 1, Card 2, and Card 3) with different alignments of categories are used. Each card is designed to protect respondent privacy. Each respondent is shown one card only. One third of the sample individuals are shown each card. Based on the pattern of responses to the three cards, it is possible to estimate the proportion of respondents in the sensitive category. For example, suppose there are at least four response categories; researchers are interested in the proportion of the population belonging to Categories 1 through 4; and Category 4 (e.g., illegal immigrant) is a sensitive category. Below are three cards that could be shown to the three thirds of the people chosen for the sample.

	Card 1	Card 2	Card 3
A	Category 1	Category 2	Category 3
B	Categories 2, 3, 4	Categories 1, 3, 4	Categories 1, 2, 4
C	Other categories	Other categories	Other categories

Suppose n people respond to each card. Let n_{A1} , n_{A2} , and n_{A3} be the number of respondents who choose Category A on the three cards, respectively. Let n_{B1} , n_{B2} , and n_{B3} be the number of respondents who choose Category B on the three cards, respectively. One estimate of the proportion of people belonging to category 4 is

$$(1/3)((n_{B1} + n_{B2} + n_{B3}) - 2(n_{A1} + n_{A2} + n_{A3}))/n.$$

A related set of methods for asking about sensitive questions are called *randomized-response* techniques.

In one version of randomized response, some respondents answer a sensitive question directly, but a randomly selected subset answers a different, noncontroversial question. Privacy is protected, because the interviewer does not know whether the subject is answering the sensitive or the nonsensitive question. Statistical methods are used to estimate the proportion of the population belonging to the sensitive category. Generally, the three-card method and randomized-response methods have the advantage of eliciting more truthful responses to questions. Their disadvantage is that they take more time than a direct question to explain to a participant in a survey. There is also some loss of precision over what would be attained with a direct question. Usually, though, the desire to receive truthful information outweighs either of the disadvantages.

The method is called the three-card method because its original formulation in the context of immigration studies used three cards, with three groups per card. Depending on the application, the method can be adapted to use different numbers of cards and different numbers of groups per card. Methods of statistical design can be used to design sets of cards for a particular question that yield efficient estimates of proportions. Statistical design can also be used to optimally divide the survey respondents among the available cards, instead of performing an even split of the subjects among the cards.

—Michael D. Larsen

See also Questionnaires

Further Reading

- Droitcour, J. A., Larson, E. M., & Scheuren, F. J. (2001). The three card method: Estimating sensitive survey items—with permanent anonymity of response. *Proceedings of the American Statistical Association, Statistical Computing Section*, CD-ROM. Retrieved from <http://www.amstat.org/sections/srms/Proceedings/y2001/Proceed/00582.pdf>
- Kingsbury, N., Droitcour, J. A., & Larson, E. M. (2003). *Estimating illegal immigrants in a census or survey: The three-card method—an innovative technique*. Conference of European Statisticians, Joint ECE-Eurostat Work Session on Migration Statistics, U.N. Statistics Division,

Working paper No. 3. Retrieved from <http://www.unecce.org/stats/documents/2003/04/migration/wp.3.e.pdf>

U.S. General Accounting Office. (1999, November). *Survey methodology: An innovative technique for estimating sensitive survey items* (GAO/GGD-00-30). Washington, DC: Author. Retrieved from <http://www.gao.gov/new.items/gg00030.pdf>

THURSTONE SCALES

Thurstone scales are a set of procedures used to construct attitude scales. These procedures were among the earliest systematic methods devised to measure attitudes. In the late 1920s, Louis L. Thurstone developed his *law of comparative judgment*. This provided a foundation for the idea that attitudes lie along a psychological continuum and can be measured despite lacking a physical dimension. This allowed for quantitative investigations of subjective psychological phenomena, such as attitudes. Thurstone developed three different attitude-scaling techniques: the method of equal-appearing intervals, the method of paired comparisons, and the method of successive intervals.

In his 1928 paper, "Attitudes Can Be Measured," Thurstone described his conceptualization of attitudes as complex and multifaceted. Specifically, he thought of an attitude as having a distribution of values on a continuum, rather than having a single value, and therefore it would not be adequately represented by a single number. He defined *opinion* as a verbal expression of a person's overall attitude. Therefore, he posited that opinions can be used to measure attitudes. Based on these assumptions, Thurstone developed his three methods to measure attitudes. However, the method of equal-appearing intervals has been the most widely used of the three methods and is the method generally referred to by the term *Thurstone scales*.

The Method of Equal-Appearing Intervals

In creating a scale using the *method of equal-appearing intervals*, the first step is in determining the

attitude object that will be evaluated. The next step is to construct a set of statements about the attitude of object that captures an entire range of opinions, from extremely unfavorable (e.g., "Abortion weakens the moral fiber of our society") to neutral (e.g., "Abortion brings happiness to some, unhappiness to others") to extremely favorable (e.g., "Abortion should be encouraged for unwanted pregnancies"). Thurstone used 130 statements to develop one of his original scales, although subsequent researchers have suggested that 40 to 50 statements are typically enough to fully capture the continuum of a particular attitude object.

Once the researcher develops a complete set of statements, judges rate favorability values for each statement using an 11-point scale, which may take various forms (see Figure 1). Typically, only the two extreme ends and the midpoint of the scale are labeled. These three forms are alternatives to Thurstone's original procedure of sorting the statements, presented one at a time on cards, into 11 physical piles. Thurstone used 300 judges to obtain favorability values, but other researchers have successfully constructed scales using as few as 10 to 15 judges.

When making these ratings for the pool of statements, judges are instructed to respond objectively on the favorableness or unfavorableness of the statements, not based on their personal agreement or disagreement with the statement. Judges are also asked to treat each interval as being equal. To ensure that judges' ratings are not careless and conform to the instructions, Thurstone recommended a screening procedure to eliminate the data set of judges who placed 30 or more statements (of the 130 total) into just 1 category out of the 11. This is equivalent to placing 23% or more statements into a single category, which may be problematic considering the care taken to develop the pool of statements to capture the entire range of opinions from favorable to unfavorable. After this screening procedure, a measure of central tendency for each statement and a measure of variability are calculated across the remaining judges. Thurstone used the median and interquartile range, but in recent years, means and standard deviations

1. For each statement, circle the number that indicates the degree of favorableness of the statement toward the attitude object.

Extremely unfavorable					Neutral				Extremely favorable		
1	2	3	4	5	6	7	8	9	10	11	
1	2	3	4	5	6	7	8	9	10	11	a. Abortion weakens the moral fiber of our society.
1	2	3	4	5	6	7	8	9	10	11	b. Adoption is a good alternative to abortion.

2. For each statement, circle the letter that indicates the degree of favorableness of the statement toward the attitude object.

Extremely unfavorable					Neutral				Extremely favorable		
A	B	C	D	E	F	G	H	I	J	K	
A	B	C	D	E	F	G	H	I	J	K	a. Abortion weakens the moral fiber of our society.
A	B	C	D	E	F	G	H	I	J	K	b. Adoption is a good alternative to abortion.

3. For each statement, place a check mark on the adjacent line indicating the favorableness of the statement toward the attitude object.

Extremely unfavorable					Neutral				Extremely favorable		
-----					-----				-----		
-----					-----				-----		a. Abortion weakens the moral fiber of our society.
-----					-----				-----		b. Adoption is a good alternative to abortion.

Figure 1 Three Possible Forms of the Scale Used by Judges to Rate Favorability of Attitude Statements

have more commonly been used. Low variability indicates relative consensus among the judges, whereas high variability indicates little agreement between the judges, signifying that the particular statement may be ambiguous.

At this point, the researcher now has mean favorability values and scores of variability between judges for each statement. For example, the statement "Abortion weakens the moral fiber of our society" may have a mean favorability score of 1.6, with a standard deviation of .3. From this original pool of statements, the researcher selects 20 to 25 statements using the mean and variability scores. Optimally, the selected statements should be equally spaced across the entire range. For developing a 20-item scale, items should be selected at intervals of .5. (There are 10 units from 1 to 11 on an 11-point scale; $10 \text{ units} \div 20 \text{ items} = .5$.) Therefore, if the researcher were using the example statement with a mean value of 1.6, he or she would need to select 19 other statements with intervals of .5 between statements (i.e., a statement closest to each of the following values: 1.1, 1.6, 2.1, 2.6, 3.1, 3.6 . . . up to 10.6). These selected items should also have low variability ratings across judges.

Once the researcher selects 20 to 25 statements from the original set of statements using these criteria, these selected statements are presented, randomized, in a checklist form to create the final scale. In Figure 2, mean favorability values have been reported in

parentheses following each statement, to illustrate that the statements are presented in random order and show the value associated with the varying degrees of opinions. However, these values would not appear in an actual scale given to participants. Participants are instructed to place a check mark beside any of the statements with which they agree. The scale score for each respondent is calculated by finding the mean or median of the total scale values that they agree with. For example, a participant may agree with a total of three statements in Figure 2, Items 4, 10, and 17. The corresponding scale values are 6.0, 8.7, and 7.6. The participant's overall scale score is obtained by taking the mean value, 7.4 in this case. A score of 7.4 indicates that this participant has a moderately favorable attitude toward abortion.

The Method of Paired Comparisons

In addition to the method of equal-appearing intervals, Thurstone also developed two other less commonly used methods. In the *method of paired comparisons*, each attitude statement is paired with every other attitude statement. For each pair, judges decide which statement of the pair is more favorable toward the attitude object. This information is used to position each statement on the attitude continuum. In this method, as in the other two methods, the final scale is constructed by eliminating ambiguous items, marked by high variability

Check (✓) the statements with which you agree.

- ___ 1. Abortion brings happiness to some, unhappiness to others. (5.1)
- ___ 2. Abortion should be encouraged for all unwanted pregnancies. (10.4)
- ___ 3. Abortion prevents an unwanted child growing up in an orphanage. (8.0)
- ___ 4. Abortion is allowable if there is no financial cost to the public. (6.0)
- ___ 5. Abortion weakens the moral fiber of our society. (1.6)
- ___ 6. We must allow abortions for certain situations. (8.7)
- ___ 7. Abortion is never justified. (1.1)
- ___ 8. It doesn't make any difference to me whether we allow abortions or not. (5.6)
- ___ 9. Most people are too lax about allowing abortions. (3.2)
- ___ 10. People who get abortions deserve our understanding, but not special treatment. (7.6)
- ___ 11. Abortions should be easier to obtain than they are now. (9.1)
- ___ 12. Abortion is justified only in instances of rape. (3.6)
- ___ 13. We can't call ourselves civilized as long as we allow abortions. (2.5)
- ___ 14. Abortion may be unacceptable, but people deserve our sympathy. (4.6)
- ___ 15. If a woman wants an abortion, it is her right. (10.1)
- ___ 16. Abortion may be the best option in certain cases. (7.1)
- ___ 17. People who get abortions should be treated like everybody else. (9.5)
- ___ 18. Abortion cannot be regarded as a sane method of dealing with unwanted pregnancies. (2.2)
- ___ 19. I think allowing abortion is necessary, but I wish it weren't. (6.6)
- ___ 20. Giving up a child for adoption is a good alternative to abortion. (4.1)

Figure 2 Opinions About Abortion

between the judges, as well as eliminating irrelevant items that fail to discriminate people with differing attitudes from the pool of scaled attitude statements. A limitation of this method stems from the need to compare each statement with every other statement. For example, 10 statements require judging 45 pairings, and 20 statements require 190 pairings, making this technique cumbersome and unmanageable for scaling a large number of attitude statements.

The Method of Successive Intervals

Finally, the *method of successive intervals* is an extension of the method of equal-appearing intervals. As in the method of equal-appearing intervals, raters are asked to make one comparative judgment for each attitude statement, placing statements in categories

of varying degrees of favorability. However, in this method, equal intervals are determined statistically, rather than relying on raters' subjective judgments of equal intervals. Specifically, the interval widths are obtained based on the assumption that the distribution of the judgments follows the normal curve. Estimates of the widths of the intervals as well as the boundaries relative to a statement's scale value are obtained from the proportion of raters who place a statement in each favorability category. The final scale is constructed in the same way as the other two methods.

Evaluating Thurstone Scales

Reliability for the scale is high with Cronbach's alpha, typically reported in the .80s. The scale also has high test-retest reliability, in the range of .90 to .95. Compared with other scaling methods, it is considerably easier using the Thurstone method to create alternative forms of the same scale. This is because the researcher has a pool of statements that have been scaled by judges. To construct a second equivalent scale, the experimenter may simply choose statements from the pool that are comparable to, but not chosen for, the first complete scale. Thurstone scales also have the advantage of having a clear neutral point on the scale, unlike the Likert scale. This neutral point allows for absolute interpretation of scale scores.

One main disadvantage of Thurstone scales is that they are labor-intensive. They require generation of a comprehensive pool of sample statements, scaling the favorability of the statements using at least 10 to 15 judges, then selecting 20 to 25 of the statements to construct a final scale. Also, in the method of equal-appearing intervals, the researcher assumes that judges will follow the instruction of treating the intervals as being equal, but this assumption cannot be tested to verify whether it is warranted. Finally, there has been criticism regarding whether judges' personal attitudes affect the perceived favorability of the attitude statements. Although the evidence is mixed in this regard, this problem may affect the scaling of statements for the method of equal-appearing intervals, but the problem has shown to have little or no effect

for the method of successive intervals and the method of paired comparisons.

Although Thurstone scales were one of the first applications of formal scaling techniques to attitude measurement, they are still accepted as a theoretically sound and valid method of attitude measurement today. Thurstone scales are still in use. However, they are not the most commonly used form of attitude measurement, as some other less labor-intensive methods can produce attitude scales with adequate psychometric properties.

—Leandre R. Fabrigar and J. Shelly Paik

See also Likert Scaling

Further Reading

- Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
- Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Fort Worth, TX: Harcourt Brace Jovanovich.
- Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.
- Ostrom, T. M. (1989). Interdependence of attitude theory and measurement. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 11–36). Hillsdale, NJ: Erlbaum.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Marsella, A. J., Dubanoski, J., Hamada, W. C., & Morse, H. (2000). The measurement of personality across cultures: Historical, conceptual, and methodological issues and considerations. *American Behavioral Scientist*, 44(1), 41–62.

This article discusses historical, conceptual, and methodological issues (including familiarity with different types of measurement tools, such as Likert and **Thurstone Scales**) associated with the cross-cultural measurement of personality. It documents the extensive debate and discussion that have emerged from the juxtaposition of the trait-situation, universalism-relativism, quantitative-qualitative, and anthropology-psychology polarities

in the past decade. Following a discussion of these polarities, the article concludes that the contending (and contentious) positions should be replaced by collaborative disciplinary research efforts that are open to the possibility of both cultural variations and universals in human behavior. Fundamental similarities in behavior may exist across cultural boundaries because of bioevolutionary, natural language descriptors, and similar life activity and socialization contexts, and major differences may exist for the very same reasons.

TIME SERIES ANALYSIS

A time series is an ordered sequence of observations over time, and very often there is dependence between the past and the future values of these observations. Time series analysis is the branch of statistics that makes use of this serial dependence for the purpose of modeling.

An effective way to graphically represent a time series is by putting the observations on a Cartesian plane, with the times of occurrence as abscissas and the values of the observations as ordinates. Such a graphical representation is called a time plot. Figure 1, representing the quarterly time series of food retail sales in New Zealand (in millions of New Zealand dollars), provides an example.

Structural Models

Typically in a time series, four components can be distinguished: trend (*T*), cycle (*C*), seasonality (*S*), and irregular component (*I*). The *trend* is the longest-term behavior of the time series, and the *cycle* is a long-term cyclical component. Sometimes the time series is not long enough to distinguish between trend and cycle, and so a unique trend-cycle component can be considered. The *seasonality* is a short-term cyclical component often due to the seasons, and, finally, the *irregular component* is an erratic component obtained as residual once the other components are identified and removed. The nonsystematic behavior of the

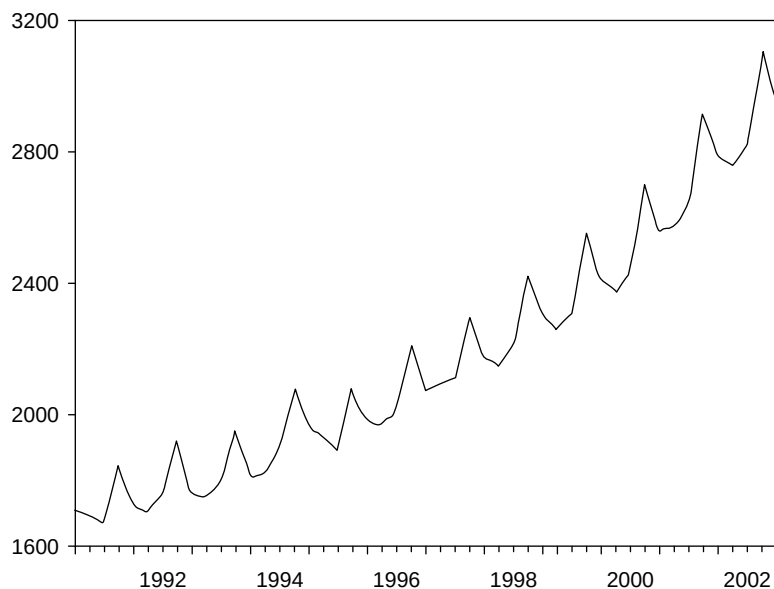


Figure 1 Food Retail Sales in New Zealand

irregular component is used to assess the goodness of the decomposition, and often it can be reasonably assumed that it follows a normal distribution.

There are two main basic ways in which the components mix together: the additive model

$$X = T + C + S + I, \quad (1)$$

where X is the original time series, and the multiplicative model

$$X = (T)(C)(S)(I).$$

If the oscillations of the cyclical components increase with a higher level of the time series, a multiplicative model is more appropriate; if the oscillations remain constant, an additive model is preferable.

A possible way to find the different components is based on regression, and the time series of food retail sales provides an example to illustrate it.

Food retail sales have a tendency to increase over time, and this could be explained by the presence of a trend or by an ascending part of a cycle, although given the nature of the series, affected by price inflation, it is more likely to be a trend. We then consider three components

in this decomposition: trend-cycle, seasonality, and irregular component. The oscillations of the seasonal appear independent from the level of the series, suggesting an additive model.

It is a good strategy to remove the longer-term components first, and in this case, the trend-cycle ($T + C$) is exponential and can be estimated by the equation

$$X = c + \alpha t + \beta t^2 + \varepsilon,$$

where t is a regular ascending sequence (0,1,2,...) representing a counter for the time and ε is the residual part. Consequently, the trend-cycle component is given by

$$T + C = c + \alpha t + \beta t^2,$$

where c , α , and β are the parameters to estimate. More specifically, the estimated equation by least squares is

$$T + C = 1729.54 + 4.88t + 0.43t^2.$$

The trend-cycle component estimated by the equation above is shown in Figure 2.

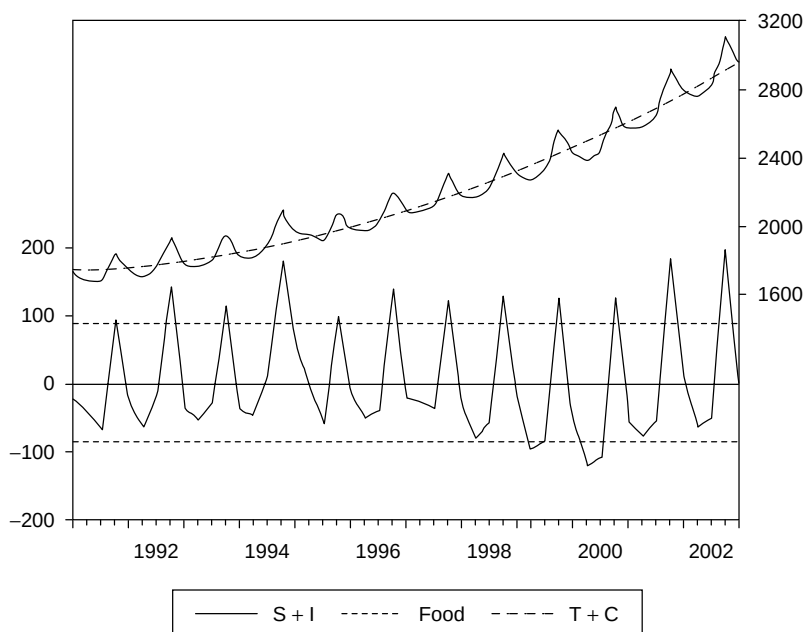


Figure 2 Trend-Cycle, Food Retail Sales, and Seasonal and Irregular Components

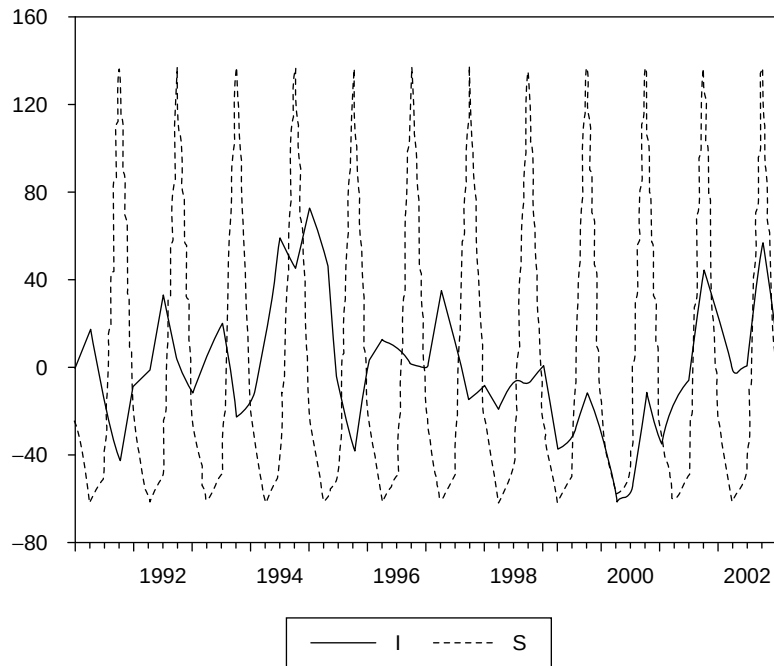


Figure 3 Irregular and Seasonal Components

The estimated component should then be removed from the original time series, and following Equation 1, we have

$$X - (T + C) = S + I. \quad (2)$$

The seasonal component can be identified by the equation

$$S + I = f_1 + f_2 + f_3 + f_4 + \varepsilon,$$

where the f s are the seasonal factors, one for each quarter, as the series is quarterly. The seasonal component is estimated by the equation

$$S = f_1 + f_2 + f_3 + f_4$$

with a least squares estimate given by

$$S = -23.84f_1 - 61.80f_2 - 49.39f_3 + 137f_4,$$

where the sum of the weights is not far away from zero, the theoretical value. The irregular component is then found as a residual by the equation

$$I = (S + I) - S.$$

Eventually, the goodness of the decomposition can be assessed by an inspection of the distribution of the

irregular component. For the example of the food retail sales, the distribution of the irregular components is approximated by the histogram in Figure 4, which is close to a Normal.

ARIMA Models

A simple way to model the serial correlation in a time series is by expressing the current observation (x_t) as a function of the past observations ($x_{t-1}, x_{t-2}, \dots, x_{t-p}$) in a regression equation,

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t, \quad (3)$$

where

α is a constant that can be omitted if the time series is mean corrected,

$\phi_1, \phi_2, \dots, \phi_p$ are the parameters associated with the past observations, and

ε_t is the regression error, which should have zero mean and be uncorrelated with previous and future regression errors.

It is also often desirable for estimation and testing purposes that the regression error have a Normal distribution.

Equation 3 describes a regression of a variable on its own past, and for this reason is called *autoregression*;

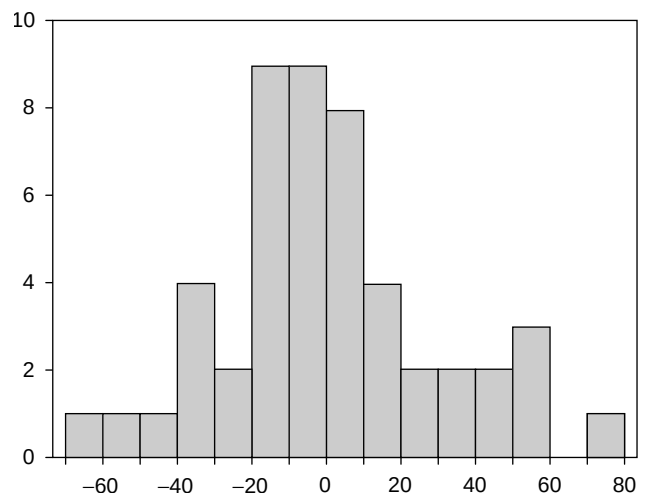


Figure 4 Distribution of the Irregular Component

in particular, it is an autoregression of order p , indicated as $AR(p)$, because we consider observations up to x_{t-p} . A desirable feature for a time series model is stationarity with respect to the mean, where we would like a particular estimated time series model to provide an adequate estimation of the mean of x_t regardless of the specific time t considered. If a trend is present, the time series is not stationary, as its expected value depends on t ; nevertheless, the difference of successive values of the observations is an effective way to remove the trend.

Consider, as an example, a time series with an additive deterministic trend modeled by the equation

$$x_t = \alpha + \beta t + \varepsilon_t, \quad (4)$$

where

α is a constant,

β is the slope of the trend t , and

ε_t is a zero mean stationary component.

Its expected value,

$$E(x_t) = \alpha + \beta t,$$

depends on t . If we consider the first difference,

$$\begin{aligned} \Delta x_t &= x_t - x_{t-1} = \alpha + \beta t + z_t \\ &\quad - [\alpha + \beta(t-1) + z_{t-1}] \\ &= \beta + \varepsilon_t - \varepsilon_{t-1}, \end{aligned}$$

its expected value,

$$E(\Delta x_t) = \beta,$$

does not depend on t any more, the expected values of ε_t and ε_{t-1} being identical, given the stationarity in mean of ε_t .

Sometimes a higher order is needed to obtain stationarity in mean. As an example, consider the case of a time series with a second-order polynomial additive trend, modeled by the equation

$$x_t = \alpha + \beta_1 t + \beta_2 t^2 + \varepsilon_t. \quad (5)$$

Its first difference,

$$\begin{aligned} \Delta x_t &= x_t - x_{t-1} = \alpha + \beta_1 t + \beta_2 t^2 + \varepsilon_t \\ &\quad - [\alpha + \beta_1(t-1) + \beta_2(t-1)^2 + \varepsilon_{t-1}] \\ &= \beta_1 + 2\beta_2 t - \beta_2 + \varepsilon_t - \varepsilon_{t-1}, \end{aligned}$$

has expected value

$$E(\Delta x_t) = \beta_1 + 2\beta_2 t - \beta_2,$$

which still depends on t and hence is not stationary in mean; nevertheless, the second-order difference,

$$\begin{aligned} \Delta^2 x_t &= (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \\ &= 2\beta_2 + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}, \end{aligned}$$

has expected value

$$E(\Delta^2 x_t) = 2\beta_2,$$

which is stationary in mean.

The order, d , of the difference we need to consider to obtain stationarity is called *order of integration of the time series*, and it is indicated as $I(d)$; the time series arising from Equation 4 has order of integration one, $I(1)$, and the time series described by Equation 5 has order of integration two, $I(2)$.

In principle, we could consider any order of integration, but in practice, it is sensible to consider up to second order, as higher orders of integrations corresponding to polynomial trends of order higher than 2 may describe stationary cyclic time series.

An $AR(p)$ model can also be represented as a moving average of the errors ε_t . Consider the simple case of an $AR(1)$ model,

$$x_t = \phi x_{t-1} + \varepsilon_t.$$

By solving recursively,

$$\begin{aligned} x_t &= \phi(\phi x_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \dots = \phi^n x_{t-n} + \sum_{i=0}^{n-1} \phi^i \varepsilon_{t-i}, \end{aligned}$$

and for a large n and $|\phi| < 1$,

$$x_t \approx \sum_i \phi^i \varepsilon_{t-i}.$$

This suggests the possibility of modeling a time series by using a moving average of the errors ε_{t-i} , $i = 0, 1, \dots, q$; that is,

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}, \quad (6)$$

where the parameters θ_i can be negative. Such a model is called a *moving average of order q* and is indicated as $MA(q)$. In many situations, the most parsimonious model in terms of parameters is obtained with a mixed model containing both AR and MA components. This leads to the general $ARIMA(p, d, q)$ model; that is, after having made a time series stationary by differencing it d times, we can show the model as follows:

$$\begin{aligned} x_t - \phi_1 x_{t-1} - \dots - \phi_p x_{t-p} &= \varepsilon_t \\ - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}. \end{aligned} \quad (7)$$

Without any loss of generality, Equation 7 assumes the stationary time series x_t to have zero mean; this can easily be obtained by subtracting a nonzero mean from its original time series. Equation 7 encompasses as special cases the pure $AR(p)$ and $MA(q)$ models.

The orders p and q of AR and MA models can be identified by inspecting, respectively, the partial autocorrelation function and the autocorrelation function. An $AR(p)$ model is adequate if the autocorrelation function is not significantly different from zero after the lag p . Similarly, a $MA(q)$ model is adequate if the autocorrelation function is not significantly different from zero after the lag q . Both conditions ensure the adequacy of an $ARMA(p, q)$ model.

$ARMA$ models are for single time series, but sometimes the consideration of more time series jointly is preferable so that their interdependence can also be modeled. If this is the case, $ARMA$ models can be generalized to a multivariate situation. Such models are called *vector ARMA models* or, more simply, *VARMA models*.

Using the Computer

For the analysis of food retail sales, the software EVIEWS was used. Another suggested software is R, which can be freely downloaded from the Internet.

—Marco Reale

See also Autocorrelation; Fourier Transform; Part and Partial Correlations; Regression Analysis

Further Reading

- Chatfield, C. (2003). *The analysis of time series: An introduction*. (6th ed.). Toronto, Canada: Chapman & Hall.
- Ghysels, E., & Osborn, D. R. (2001). *The econometric analysis of seasonal time series*. New York: Cambridge University Press.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Lutkepohl, H. (2005). *New introduction to multiple time series*. New York: Springer.
- Pena, D., Tiao, G. C., & Tsay, R. S. (2000). *A course in time series analysis*. New York: Wiley.
- Shumway, R. H., & Stoffer, D. S. (2005). *Time series analysis and its applications*. New York: Springer.

R software and manuals: <http://stat.cmu.edu/R/CRAN/>
Time Series Data Library: <http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/>

TORRANCE, E. PAUL (1915–2003)

Ellis Paul Torrance is known throughout the world for his eponymous tests of creative thinking. Yet he considered the tests a means to an end rather than an end in themselves. His greatest interest and deepest sense of pride came from his work in developing creativity in all people.

Born on a farm in Milledgeville, Georgia, Paul was physically unable to plow a straight line or pick much cotton, but he showed precocious intellectual abilities. Thus, his father urged him to pursue an education. Lacking the funds to attend college during the Depression, Paul worked and saved to attend school when and where he could, including taking some classes by correspondence and enrolling one summer

as the only male at the Georgia State College for Women. Afterward, teaching at Georgia Military College, he became very interested in psychology and the students labeled as “troublemakers” who had a special “spark,” so he went on to earn an MA from the University of Minnesota. Lacking the funds to pursue the doctorate, he returned to a teaching and counseling career until he was drafted into the army to serve as a psychologist. There, he saw the same spark in the fighter pilots that he trained for survival and concluded that the spark was related to creativity. After completing his term of duty, he received the financial support needed to obtain his doctorate at the University of Michigan, in 1951.

In 1958, the launch of *Sputnik* and Guilford’s presidential speech to the American Psychological Association (APA) on creativity spurred the research push that provided Torrance the opportunity to study creativity. He was hired as the director of the Bureau of Educational Research at the University of Minnesota and commissioned to begin a 25-year study on giftedness. Torrance’s studies of creativity extended for more than 40 years.

As part of the research, Torrance developed many creativity measures, with the Torrance Tests of Creative Thinking (TTCT) and Thinking Creatively in Action and Movement (TCAM) being two of the best known. His studies with these measures helped dispel the idea that IQ tests alone gauged all intelligence. His 7-year, 12-year, 22-year, and 40-year longitudinal studies showed a strong relationship between test behavior in childhood and adult real-life creative behavior, thereby offering evidence of the predictive validity of the TTCT.

However, Torrance was most proud of the Future Problem Solving Program (FPSP) and the Incubation Model of Teaching, which he created to infuse creativity and relevancy into the school curriculum. Started in 1974, the FPSP is now an international program, with over 250,000 student participants and over 40 affiliate programs. The Incubation Model of Teaching, a three-dimensional curriculum model with the goal of encouraging children’s inherent curiosity, includes methods for guiding students to think about what they have learned, delve into the topic by asking

questions and experimenting, and, ultimately, use what they have learned. Upon Torrance’s retirement in 1984, The Torrance Center for Creative Studies was established to continue scholarly inquiry into the study of creativity.

—Bonnie Cramond and Kyung Hee Kim

See also Torrance Tests of Creative Thinking; Torrance Thinking Creatively in Action and Movement

Further Reading

Millar, G. W. (1995). *The creativity man: An authorized biography*. Norwood, NJ: Ablex.

E. Paul Torrance: <http://www.coe.uga.edu/torrance/>

TORRANCE TESTS OF CREATIVE THINKING

The Torrance Tests of Creative Thinking (TTCT, published by Scholastic Testing Service, <http://www.ststesting.com>) are a battery of figural and verbal tests designed to assess creative thinking in individuals from kindergarten through adulthood. Just as IQ tests are measures of some cognitive abilities related to intelligence, the TTCT are measures of intellectual abilities that are used in creative achievements. They are not purported to measure the entire global construct of creativity, as they don’t measure motivation, skill, or any other of the many components that may impact an individual’s ultimate creative productivity. Yet in the 40-plus years since they were created, the tests have been translated into more than 35 languages and have been used around the world. There are several studies affirming their predictive validity, most recently the results of the 40-year follow-up of elementary children who were given the tests in 1958 and were contacted in 1998 to assess their creative achievements in adulthood.

Torrance believed that everyone has creativity and that it can be nurtured. He designed the tests to measure creative thinking abilities so that they could be enhanced. The tests were seen as a means of assessing the effectiveness of creativity training, pointing

out abilities that might otherwise go unnoticed, understanding the human mind, and assisting with curriculum design and psychotherapy. He and his students and colleagues tested many stimuli to find ones that were motivating to children and adults, gender neutral, and not discriminatory to children from different racial or socioeconomic groups. The resultant tests are used for all ages, although the norms are age- or grade-based for conversion to standard scores.

The verbal tests consist of six activities and take about 1 hour to administer. The respondents are requested to ask questions, guess causes, predict consequences, improve a product, think of new uses for a common object, and reason in a hypothetical situation. The figural tests consist of three activities and take approximately 45 minutes to administer. The respondents are given 10 minutes each to add details to black-and-white shapes and abstract line drawings to make something out of them. The instructions for the activities are designed to motivate the respondents to give creative responses by instructing them to give many unusual, detailed ideas.

The TTCT are most often used as part of a multiple-criterion approach to identifying students for gifted programs. The recent and growing emphasis on identifying a broad array of talents in a diverse population of students has increased interest in assessments like the TTCT, especially the figural forms, which are not heavily dependent on language. Also, because Torrance was originally interested in creative students, “wild colts” who were often in trouble in schools, the TTCT may be particularly useful for discovering and redirecting such children’s energies and talents toward more positive pursuits.

—Bonnie Cramond

See also Torrance, E. Paul; Torrance Thinking Creatively in Action and Movement

Further Reading

Center for Creative Learning. (n.d.). Creativity assessment test number 72. Retrieved August 30, 2005, from <http://www.creativelearning.com/Assess/test72.htm>

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. *Gifted Child Quarterly*, 49, 283–291.

Torrance, E. P. (1967/1995). Rationale of creativity tests. In E. P. Torrance (Ed.), *Why fly? A philosophy of creativity* (pp. 83–99). Norwood, NJ: Ablex.

Torrance, E. P., & Goff, K. (2001). *Brief demonstrator Torrance Tests of Creative Thinking*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P., & Safter, H. T. (1999). *Making the creative leap beyond*. . . . Buffalo, NY: Creative Education Press.

TORRANCE THINKING CREATIVELY IN ACTION AND MOVEMENT

Thinking Creatively in Action and Movement (TCAM) was developed by E. Paul Torrance in 1981. Scholastic Testing Service, Inc. holds the copyright. The TCAM is designed to measure fluency, originality, and imagination in children age 3 to 8 through various movement and manipulation exercises.

The test consists of four activities:

Activity I, “How Many Ways?” is used to observe the child’s ability to move in alternate ways across the floor.

Activity II, “Can You Move Like?” asks the child to move like animals or a tree.

Activity III, “What Other Ways?” has the child place a paper cup in a wastebasket in alternate ways.

Activity IV, “What Might It Be?” involves the child coming up with a variety of uses for a paper cup.

The activities use paper cups, a wastebasket, pencils, and strips of red and yellow tape. The TCAM is administered individually, and it takes about 15 minutes. There is no time limit, but the administrator should keep a record of the time used. It is recommended that time be limited to a period that will not overly fatigue the child (generally 10–30 minutes). Only one child should be in the activity room, with enough space for movement. Warm-up and motivational procedures are suggested before administration so that children can relax and have fun with the activities.

The examiner records the variety of verbal and kinetic responses. A clear and easy-to-follow scoring guide is provided, with directions for scoring and administration procedures. Activities I, III, and IV are scored for fluency and originality. Activity II is scored for imagination. Fluency scores are the number of relevant responses; originality scores range from 0 to 3 points for each response and are based on an originality list derived from the statistical frequency of responses; and imagination scores are based on a 5-point Likert scale ranging from “no movement” to “excellent; like the thing.”

Norms for the TCAM are based on 1,896 children, age 3 to 8 years old, from 11 states and Guam, with equal representation among Whites and Blacks. Tables for converting raw scores to standard scores ($M = 100$; $SD = 20$) are reported only for 3- to 6-year-olds.

Interscorer reliability coefficients were reported as .90 to .99. Test-retest reliability coefficients were .84 for a sample of twenty 3- to 5-year-olds for a 2-week interval, and .71 to .89 for a sample of thirty 7- to 8-year-old boys with learning disabilities for 1- to 14-day intervals. In this study, the alpha coefficient for the test internal consistency was .79.

Validity studies showed significant positive correlations between TCAM and other characteristics of creativity, such as a modified Piaget measure of divergent thinking, a divergent problem-solving-based mathematics test, the production of humor, and the Multidimensional Stimulus Fluency Measure.

There are many advantages of using the TCAM. Studies found no evidence of bias for sex, socioeconomic status, or race. Another advantage is that the TCAM can be used with children with impairments, such as those who are emotionally disturbed or deaf and those who have behavior disorders or physical disabilities. Furthermore, the use of movement is an appropriate means of reaching young children's creativity because it makes sense to children and can be administered in a playful, gamelike environment, rather than a sterile, testlike situation. Last, the TCAM is useful for teaching creative movement and brainstorming techniques, which can foster creativity in young children. One concern, however, is that the TCAM has not been renormed nor have the originality lists been updated since 1981. Because of this, the

credibility of the norms and originality scores can be questioned, because responses may have changed between 1981 and the present.

E. Paul Torrance is best known for developing the Torrance Tests of Creative Thinking (TTCT), which are the most widely used tests of creativity, for kindergarteners through adults.

—Kyung Hee Kim

See also Torrance, E. Paul; Torrance Tests of Creative Thinking

TREE DIAGRAM

A tree diagram is used to summarize the probabilities associated with a sequence of random events. The set of branches emanating from any given start point represent all the possible events that could follow. Each branch is labeled according to the probability of the event occurring, given the events that have previously happened. Hence, the sum of probabilities from each set of branches must equal 1.

Each path from the start of the tree to the end defines an outcome in the sample space. The outcomes defined by the paths are mutually exclusive. All outcomes in the sample space are represented by a path. The probability of each outcome is obtained by multiplying the conditional probabilities along the path, which is often called the “Multiplication Rule.” Therefore, these probabilities sum to 1.

An interesting application of tree diagrams is provided by the following example, toward anonymizing a survey question. Consider the following:

- Toss a fair coin.
- If you get a “head,” answer the question:
 - Have you ever cheated in an exam? Yes or No.
- If you get a “tail,” answer the question:
 - Flip coin a second time. Did you get a head on the second flip? Yes or No.

It is not possible to tell whether a particular individual answered “Yes” because they cheated in an exam or got a head on the second flip. However, if we have a large sample of responses, we can estimate the proportion of people who have cheated in an exam.

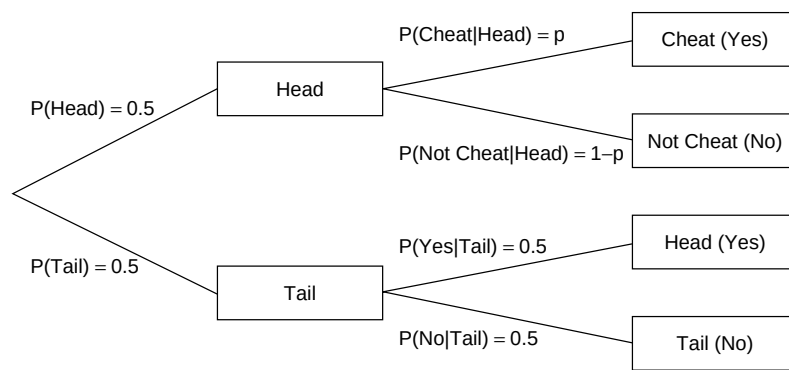


Figure 1 Tree Diagram for Anonymized Exam Cheating Question

Consider the tree diagram in Figure 1, created in PowerPoint.

The probability of answering “Yes” is given by

$$\begin{aligned}
 P(\text{Yes}) &= P(\text{Head} \cap \text{Cheat}) && \text{(addition across} \\
 &+ P(\text{Tail} \cap \text{Head}) && \text{branches)} \\
 &= 0.5 \times p + 0.5 \times 0.5 && \text{(multiplying} \\
 & && \text{along path)} \\
 &= 0.5p + 0.25
 \end{aligned}$$

This calculation assumes a large sample with fair coins to minimize bias in the coin flip probabilities. If 35% of sample answered “Yes,” then

$$0.35 = 0.5p + 0.25,$$

which can be rearranged to give

$$p = P(\text{Cheat}|\text{Head}) = 2 \times (0.35 - 0.25) = 0.2.$$

The probability that someone cheated in an exam is independent of the first head, so $P(\text{Cheat}|\text{Head}) = P(\text{Cheat})$. Hence, from the survey, we estimate that around 20% of students have cheated in an exam.

—Carl J. Scarrott

See also Conditional Probability

Further Reading

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: W. H. Freeman.

Drawing tree diagrams: <http://www-stat.stanford.edu/~susan/surprise/ProbabilityTree.html>

TRUE/FALSE ITEMS

True/false items are used for achievement type tests when there is a clear distinction between the two alternatives, true and false. One of the best criteria for judging the value of a true/false item is whether the correct answer (be it true or false) is unequivocally the right one, the only one, and the correct one.

There are several things to keep in mind when writing true/false items:

1. True/false items are always stated as declarative sentences.
2. The alternative answers can be true-false, right-wrong, yes-no, like-dislike, and so on—as long as they are very clear choices.
3. A good true/false item focuses on one and only one idea, concept, or specific topic. With too many ideas, the test taker can become confused.
4. Double negatives should not be used in true/false questions.
5. Clues to the answer to a true/false item should not be included in the item.
6. A different type of item should be used for more complex types of inquiries that assess higher-order thinking.
7. It is preferable to have equal number of true and false items on the test. This minimizes the role that chance plays.

The Advantages and Disadvantages of True/False Questions

The advantages of multiple-choice items are considerable. First, they are convenient, such that several can be administered in a short amount of time. Second, they are very easy to score. If well written, the answers are either correct or incorrect.

But, as with all items, there are disadvantages as well. First, true/false items place a premium on memorization. It is tough to get beyond the most basic levels of knowledge with true/false items. Second, it is relatively easy to guess correctly; the probability of being right (or wrong) is 50%, and by chance alone, if the test taker selects T or F on a somewhat random

basis, the final score will be about 50%. Still, the odds of guessing correctly are much higher than in any other type of traditional item.

True/false test scores can be corrected for guessing (that is the chance outcome of getting 50% correct by guessing alone). This is a useful adjustment to make so that scores more truly reflect those who really do know more versus those who just guess.

Let's propose a scoring system where students get one point for being correct, one point for being incorrect, and nothing for leaving the item blank. So, the formula for correction (or CS, for corrected score) becomes

$$CS = R - W,$$

where

CS is the corrected score,

R is the number correct, and

W is the number incorrect.

For example, take the example of a 50-item test, where you would expect a score of 25 by chance alone ($.5 \times 50 = 25$). Bruce gets 35 correct on a 50-item test, and Bill gets 25 correct. How can we adjust these scores so that Bruce's performance (which is way above chance) is recognized?

Correcting the scores, it turns out that Bruce's new one is $35 - 15 = 20$, and Bill's is $25 - 25 = 0$. Bill is clearly "punished" for guessing.

Table 1 Corrected Scores on 50-Item Test

# Right	# Wrong	Corrected Score
50	0	50
45	5	40
40	10	30
35	15	20
30	20	10
25 (chance)	25 (chance)	0
20	30	-10
15	35	-20
10	40	-30
5	45	-40
0	50	-50

Table 1 shows the number correct on a 50-item test, the number wrong, and the corrected score.

—Neil J. Salkind

See also Ability Tests; Essay Items; Multiple-Choice Items

Further Reading

Salkind, N. J. (2006). *Tests and measurement for people who (think they) hate tests and measurements*. Thousand Oaks, CA: Sage.

True/false item test administering using Blackboard:
<http://www.mc.maricopa.edu/other/ctl/coursematerial/respondus/importfromword.pdf>

TRUE SCORE

Classical true-score theory, most often traced to the work of Charles Spearman, has been profoundly influential in educational and psychological measurement since around the turn of the 20th century. True-score theory is essentially a model of relationships between measurement errors and observed test scores. The central notion is that any observed score can be decomposed into a "true" score component and a random-error term. Importantly, different sources of error that contribute to the observed score are not differentiated within the latter term.

Classical true-score theory is expressed symbolically in the well-known expression $X = T + E$, in which X represents an observed score for a test taker, T represents that test taker's true score, and E is the error score, or the error of measurement associated with that observed score. In this model, the "true" score component of any individual's observed test score is presumed to reflect the actual amount of the characteristic being measured by the test that is possessed by the individual—that is, the score that the person would have obtained in the absence of measurement errors. Measurement errors (i.e., discrepancies between the individual's observed and true scores) can result from any number of factors both

intrinsic and extrinsic to the test taker (e.g., the specific day on which the test is taken).

For any given test taker and test, T is assumed to be fixed, while E and X are assumed to vary for that same test taker across conditions. If a student is capable of attaining a score of 15 correct from 20 questions on a particular test but is ill on the day of testing and misses 5 questions at random, the true score T for that student would be 15; the observed score X would be 10; and the error score E would be -5 . If the same student takes the test again and this time not only performs at his or her ability level but also “gets lucky” in guessing two multiple-choice questions, that student’s X would be 17; E would be $+2$; but T would still be 15.

The utility and generality of any test model hinges on the assumptions it makes, because these dictate the conditions under which the model may be assumed to hold. Classical true-score theory is, accordingly, defined by a set of assumptions that dictate the circumstances under which the model may reasonably be applied. Most of these assumptions are associated with the definition of measurement error in the model as *random* (rather than *systematic*) discrepancies between true and observed test scores:

1. True scores and error scores combine *additively* to form observed scores. That is, there are no multiplicative or other relationships between true and error scores.

2. The population mean (expected value) of observed X scores is equal to the test taker’s true score, T , that is, $\epsilon(X) = T$. Stated alternatively, if the same test taker could complete a given test an infinite number of times, the mean of the scores obtained by that test taker on that test should be equal to his or her true score. This assumption, in fact, defines T —that is, T is the mean of the theoretical distribution of observed scores that a test taker would obtain across infinite independent attempts at the same test.

3. The error and true scores within a population of test takers are uncorrelated (i.e., $\rho_{ET} = 0$). This assumption stipulates that high or low scores on a given test should not be systematically associated

with more positive or negative error scores on that test. Consider, for example, a test in which one question is unsolvable. If all of the higher-achieving students waste large amounts of time on this question and thus do more poorly than they otherwise would on the rest of the test, this could create a situation in which there was a negative correlation between true and error scores.

4. The error scores on two different tests are uncorrelated (i.e., $\rho_{E1E2} = 0$, with $E1$ being the error score for Test 1 and $E2$ being the error score for Test 2). This means that if a particular individual had a positive error score on one test, he or she should not be more or less likely to obtain a positive error score on a second test. For example, if a group of test takers’ scores on different tests were greatly influenced by factors that were common across the tests (e.g., fatigue), this assumption would not be tenable.

5. The error scores on one test are uncorrelated with true scores on another test ($\rho_{E1T2} = 0$). This assumption stipulates that the factors that cause measurement errors on one test are not related systematically to some characteristic being measured on another test.

Other assumptions of the model are associated with parallel and t -equivalent tests. Specifically, while t -equivalent tests are those that yield the same true scores with the exception of an additive constant, two tests are deemed to be parallel only if both their true scores and their error variances are equal in the population.

The true-score model forms the basis for what has become known as *classical reliability theory*. As observed scores are assumed to comprise both true and random error scores, observed score variance is, accordingly, assumed to comprise both true and random-error score variance. In classical reliability theory, the reliability coefficient represents the ratio of true to observed score variability. Stated alternatively, reliability represents the extent to which variance in observed scores can be attributed to unobservable true-score variance.

Since the true-score model was initially posed, a vast array of techniques for estimating measurement errors (and thus test reliability) have emerged. These estimates consider measurement errors that are due to inconsistencies across different forms of tests (e.g., parallel-forms reliability), across raters or markers (e.g., interrater consistency), in domain representation/sampling within tests (e.g., internal consistency), and across test occasions (e.g., test-retest reliability).

One major limitation of the classical approach to reliability, however, lies in its failure to consider the impact of different error sources on observed scores *simultaneously*. In notions of reliability derived from the true-score model, there is an implicit assumption of overlap across different error sources—that is, an assumption that these are not cumulative and that they do not interact to create additional error variance.

Generalizability (G) theory, which subsumes classical true-score theory as a special case, provides a framework for estimating the magnitude of multiple error sources simultaneously. As such, this model allows an examination of cumulative and interactive effects among different error sources. Despite the relative advantages that this approach offers, financial and other practical constraints continue to make approaches based on classical true-score theory an attractive option for test developers.

—Elaine Chapman

See also Reliability Theory; Validity Theory

Further Reading

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161–169.
- Thompson, B. (1991). Review of generalizability theory: A primer by R. J. Shavelson & N. W. Webb. *Educational and Psychological Measurement*, 51, 1069–1075.

TUKEY-KRAMER PROCEDURE

Tukey and Kramer proposed a procedure for pairwise testing of means in a one-way analysis of variance with unequal sample sizes. The procedure is routinely applied after a significant overall F test, although the F test is not required. In Tukey's honestly significant difference (HSD) procedure, a single critical difference, CD, is calculated for each pair of means. That critical difference uses critical values from the Studentized range statistic. In particular,

$$CD = q_{1-\alpha}(k, df_E) \sqrt{\frac{MS_E}{N}},$$

where

$q_{1-\alpha}(k, df_E)$ is the Studentized range statistic at Level α for k means and df_E ,

df_E is the error degrees of freedom,

MS_E is the error mean square, and

N is the common sample size.

In the Tukey-Kramer procedure, a different CD is required to evaluate the significance of the difference between each pair of means that are based on different sample sizes. The pairs are investigated independent of all other pairs. Critical differences are all based on the Studentized range distribution. The formula for the Tukey-Kramer CD testing means $\bar{X}_i$ and $\bar{X}_j$ is

$$CD = q_{1-\alpha}(k, df_E) \sqrt{\left[\frac{MS_E}{2} \left[\frac{1}{N_i} + \frac{1}{N_j} \right] \right]},$$

where

N_i is mean i ($i = 1, \dots, k$) and

N_j is mean j ($j = 1, \dots, k$), but $j \neq i$.

Anthony J. Hayter proved that the Tukey-Kramer is conservative, that is, the probability of one or more Type I error never exceeding α even if applied without a significant F test. However, if a significant F test is required, then the Tukey-Kramer CD becomes

$$CD = q_{1-\alpha}(k-1, df_E) \sqrt{\left[\frac{MS_E}{2} \left[\frac{1}{N_i} + \frac{1}{N_j} \right] \right]}.$$

The CD has the value that would be used only with $k-1$ means. The resulting Hayter-Fisher version of the Tukey-Kramer procedure will be more powerful than the original Tukey-Kramer.

Illustrative Example

Consider the following data, in which four treatment groups are being compared to a control. Lower scores indicate better performance. The MS_{WG} for these data is 3.0326, with $df_{WG} = 179$.

	Control	Treat 1	Treat 2	Treat 3	Treat 4
Mean	4.05	3.52	3.18	3.15	2.24
N	42	24	38	40	40

The analysis of variance (ANOVA) for these data would give $F = 16.48 > 2.43 = F_{.95}(4, 160) > F_{.95}(4, 179) = CV$. Therefore, we reject the full null hypothesis at the .05 level and proceed to pairwise testing.

The CD for the Control ($N = 42$) and Treatment 1 ($N = 24$) group means is obtained from

$$\begin{aligned}
 CD &= q_{1-\alpha}(k-1, df_E) \sqrt{\left[\frac{MS_E}{2} \left[\frac{1}{N_i} + \frac{1}{N_j} \right] \right]} \\
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{42} + \frac{1}{24} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0238095 + 0.0416667]} \\
 CD &= 3.68 \sqrt{1.5163[0.065476]} \\
 &= 3.68 \sqrt{0.0992815} = 3.68(0.315089) \\
 CD &= 1.1595 = 1.16.
 \end{aligned}$$

Applying the same calculations to the Control ($N = 42$) and Treatment 2 ($N = 38$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{42} + \frac{1}{38} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0238095 + 0.0263157]} \\
 CD &= 3.68 \sqrt{1.5163[0.0501253]} \\
 &= 3.68 \sqrt{0.0760050} = 3.68(0.275690) \\
 CD &= 1.0145 = 1.01.
 \end{aligned}$$

Applying the same calculations to the Control ($N = 42$) and Treatment 3 ($N = 40$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{42} + \frac{1}{40} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0238095 + 0.025]} \\
 CD &= 3.68 \sqrt{1.5163[0.0488095]} \\
 &= 3.68 \sqrt{0.0740099} = 3.68(0.2720475) \\
 CD &= 1.0113 = 1.01.
 \end{aligned}$$

Applying the same calculations to the Treatment 1 ($N = 24$) and Treatment 2 ($N = 38$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{24} + \frac{1}{38} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0416667 + 0.0263158]} \\
 CD &= 3.68 \sqrt{1.5163[0.067982]} \\
 &= 3.68 \sqrt{0.103081} = 3.68(0.32106) \\
 CD &= 1.1815 = 1.18.
 \end{aligned}$$

Applying the same calculations to the Treatment 1 ($N = 24$) and Treatment 3 ($N = 40$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{24} + \frac{1}{40} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0416667 + 0.025]} \\
 CD &= 3.68 \sqrt{1.5163[0.066667]} \\
 &= 3.68 \sqrt{0.10108667} = 3.68(0.317941) \\
 CD &= 1.1700 = 1.17.
 \end{aligned}$$

Applying the same calculations to the Treatment 2 ($N = 38$) and Treatment 3 ($N = 40$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{38} + \frac{1}{40} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.0263158 + 0.025]} \\
 CD &= 3.68 \sqrt{1.5163[0.0513158]} \\
 &= 3.68 \sqrt{0.0778101} = 3.68(0.278944) \\
 CD &= 1.0265 = 1.03.
 \end{aligned}$$

Applying the same calculations to the Treatment 3 ($N = 40$) and Treatment 4 ($N = 40$) group means produces the CD

$$\begin{aligned}
 CD &= 3.68 \sqrt{\left[\frac{3.0326}{2} \left[\frac{1}{40} + \frac{1}{40} \right] \right]} \\
 &= 3.68 \sqrt{1.5163[0.025 + 0.025]} \\
 CD &= 3.68 \sqrt{1.5163[0.05]} \\
 &= 3.68 \sqrt{0.075815} = 3.68(0.275345) \\
 CD &= 1.0132 = 1.01.
 \end{aligned}$$

Combining all the above, we have CD values for all pairs of means.

Table 1 Critical Values for Each Pair of Means

	Tr 4	Tr 3	Tr 2	Tr 1	Control
Sample Sizes, $N =$	40	40	38	24	42
Tr 4, $N = 40$	—	1.01	1.01	1.17	1.01
Tr 3, $N = 40$		—	1.03	1.17	1.01
Tr 2, $N = 38$			—	1.18	1.01
Tr 1, $N = 24$				—	1.16

Note: Tr = treatment group.

Applying the CDs to the corresponding differences between ordered means results in the following:

Table 2 Mean Differences for Pair of Means

	Tr 4	Tr 3	Tr 2	Tr 1	Control
Sample Sizes, $N =$	40	40	38	24	42
Tr 4 = 2.24	—	.91	.94	1.28*	1.81*
Tr 3 = 3.15		—	.03	.37	.90
Tr 2 = 3.18			—	.34	.87
Tr 1 = 3.52				—	.53

Note: Tr = treatment group.

* $p < .05$

The largest difference is 1.81 between Treatment 4 and the Control. This exceeds the largest CD of 1.01 and is therefore significant. The individuals in Treatment 4 have significantly lower average scores than does the Control. The two second-largest differences are 1.28 and .90. The value of 1.28 also exceeds the CD of 1.17 and is significant. Those in Treatment 4 are significantly lower than are those in Treatment 1. However, .90 is less than the CD of 1.01 and is not significant. There is no significant difference between the mean for Treatment 3 and the Control. All other differences are less than their corresponding CD values and are therefore not significant.

Most computer packages continue to use the original and slightly less powerful version of the Tukey-Kramer. However, it is a rather simple matter to apply the more powerful Hayter-Fisher version described above. If a computer package is used to apply the Tukey-Kramer, then any nonsignificant pair can be tested by applying the more powerful Hayter-Fisher version.

—Philip H. Ramsey and Patricia Ramsey

Further Reading

- Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, 12, 61–75.
- Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81, 1000–1004.
- Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods*, 7, 504–523.
- Ramsey, P. H., & Ramsey, P. P. (1990). Critical values for two multiple comparison procedures based upon the Studentized range distribution. *Journal of Educational Statistics*, 15, 341–352.

expressed in a report or journal article simply as “significant at the .05 level.”

—Neil J. Salkind

See also Significance Level; Type II Error

Further Reading

- Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

TYPE I ERROR

Statistical significance is the degree of risk that a null hypothesis will be rejected when it is actually true. This level of risk is also known as a Type I error. Traditionally, the null (a statement of equality) says that there is no difference between the two groups. Perhaps, in reality, however, there is no difference, and if the null is rejected, an error is being made. This risk is known as a Type I error.

The level of significance has certain conventional values associated with it, such as .01 and .05. For example, if the level of significance is .01, it means that on any one test of the null hypothesis, there is a 1% chance the null hypothesis will be rejected when it is true—leading to the conclusion that there is a group difference when there really is no group difference. If the level of significance is .05, it means that on any one test of the null hypothesis, there is a 5% chance it will be rejected when the null is true—leading to the conclusion that there is a group difference when there really is no group difference. Notice that the level of significance is associated with an independent test of the null, and it is not appropriate to say that “on 100 tests of the null hypothesis, I will make errors on only 5.”

In a research report, statistical significance is usually represented as $p < .05$, read as the probability of observing that the outcome is less than .05, and often

The term *Type II error* describes a particular kind of inferential mistake one can make when using data to draw a conclusion about a null hypothesis within the strategy called *hypothesis testing*. Simply stated, a Type II error is committed when a false null hypothesis is not rejected.

To understand what a Type II error is and to gain insight into the various factors that operate to increase or decrease Type II error risk, one must first be aware of the logic and steps of hypothesis testing, a statistical strategy widely used by applied researchers in many disciplines.

Hypothesis Testing

In hypothesis testing, data are collected from one or more samples and then used to make an educated guess, or inference, as to the state of affairs within the relevant population(s). In any given study, this educated guess will be focused on a specific statistical characteristic of the population(s), such as the mean (μ), the variance (σ^2), or the correlation between two measured variables (ρ). After a researcher chooses his or her population(s), the variable(s) of interest, the statistical focus, and the planned method for analyzing the study's data, that researcher can use hypothesis testing as a strategy for making the desired statistical inference.

There are six steps involved in the most basic version of hypothesis testing. They involve the following:

1. Stating a null hypothesis (H_0)
2. Stating an alternative hypothesis (H_a)
3. Selecting a level of significance (α)
4. Collecting data from the study's sample(s)
5. Determining the probability, presuming H_0 to be true, of getting sample data like those actually collected or sample data that deviate even further from what would be expected (if H_0 were true)
6. Deciding either to reject H_0 or to fail to reject H_0

An example may help clarify the way these six steps permit a researcher to engage in hypothesis testing—and how hypothesis testing may lead to a Type II error. For this example, imagine that a researcher is hired to settle a dispute between two taxpayers from New York. One believes that if all high school seniors in the state of New York were given an intelligence test, the students' mean IQ would be higher than the national average of 100. The other believes just the opposite. Also imagine that the researcher, once hired, identifies a random sample of 25 students, that each student's intelligence is measured by a trained psychologist, and that the resulting IQ scores yield a mean of 104 with a standard deviation of 15. Finally, imagine that the researcher intended from the beginning to subject the data to a one-sample t test with a level of significance equal to .05. For this example, the six steps of the hypothesis testing procedure would be as follows:

1. $H_0: \mu = 100$.
2. $H_a: \mu \neq 100$.
3. $\alpha = .05$.
4. In the sample, $n = 25$, $M = 104$, and $SD = 15$.
5. $p = 0.1949$.
6. H_0 is not rejected.

As applied to this example, the logic of hypothesis testing is straightforward. If the null hypothesis were false, a random sample should most likely produce a mean that is dissimilar to the number specified in H_0 . However, the mean IQ of the 25 students, 104, is not very inconsistent with what would be expected if H_0 were true. (With the estimated standard error of the

mean being equal to 3.0, the t test's calculated value is equal to 1.33.) Because the probability associated with the sample ($p = .19$) is larger than the selected level of significance, the researcher decides that the evidence available is not sufficiently "at odds" with H_0 to cast doubt on the null statement that the mean IQ of high school students in New York is equal to 100. Consequently, the null hypothesis is not rejected.

In this example, it is possible that the researcher's conclusion—not to reject H_0 —may have been a mistake. Such would have been the case if the unknown population mean were equal to 115, 103, 98, or any number other than 100. The null hypothesis value of 100 was selected simply because it was the "line of demarcation" between the two taxpayers who held different opinions about the actual mean IQ of all high school students in New York. It is quite conceivable that the null hypothesis is wrong, and if that is the case, not rejecting H_0 would have been the wrong thing to do. Describing this possible mistake in statistical terms, a Type II error may have been made.

When Type II Errors Can and Cannot Occur

As indicated at the outset, a Type II error is committed when a false null hypothesis is not rejected. This kind of inferential error is always a possibility when any statistical test leads to a decision not to reject the null hypothesis being evaluated. In studies dealing with correlation, a Pearson r is often tested to see if it is significantly different from a null value of 0.00. In this kind of study, a Type II error is possible if a decision is reached not to reject $H_0: \rho = 0.00$. In studies in which k groups are compared in terms of their variances, a Type II error is possible if a fail-to-reject decision is reached when evaluating $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$. In studies in which logistic regression is used to see which of several independent variables help predict status on a nominal dependent variable, a Type II error is possible if a particular variable's odds ratio is found not to differ significantly from the null value of 1.00. It is worth repeating that a Type II

error is always a possibility when any statistical test leads to a decision not to reject the test's null hypothesis.

There are two situations in which a Type II error cannot occur. On the one hand, a Type II Error cannot be committed if the null hypothesis is rejected. (In this case, a Type I error—rejecting a true H_0 —may or may not occur, but it is impossible to incorrectly fail to reject a false null hypothesis if that H_0 is rejected.) On the other hand, it is impossible to make a Type II error if the six-step hypothesis-testing procedure is turned into a four-step procedure by eliminating the third step (setting a level of significance) and final step (deciding to reject or not reject H_0). In this four-step approach to evaluating an H_0 , the size of the data-based p -value is interpreted as providing a “level of support” for the null hypothesis. The smaller the p , the lower the level of support. With no binary decision being made about H_0 , the notion of a Type II error (as well as the notion of a Type I error) does not apply.

Synonyms for Type II Error

There are different labels for this particular kind of inferential error. A Type II error is sometimes referred to as an *acceptance error*, a *false negative*, an *error of the second kind*, or a *beta error*. The first of these synonyms come into being because a decision not to reject H_0 makes it appear as if H_0 is being accepted. (Technically, it is wrong to “accept” H_0 in hypothesis testing.) The notion of a “false negative” comes from the medical field, in which diagnostic tests sometimes yield inaccurate results. An erroneous diagnosis can indicate that a patient has a disease when he or she doesn't (a false positive) or that the patient doesn't have a disease when he or she does (a false negative). The third synonym, error of the second kind, makes sense when one realizes that any inference in hypothesis testing, like a medical diagnosis, can potentially be wrong in either of two ways. (An error of the first kind, or Type I error, is committed whenever a true null hypothesis is rejected.) The final synonym, beta error, exists because a Type I error is sometimes referred to as an *alpha error*.

Type II Error as a Probability

After a study's data are collected but prior to the time a decision is reached about H_0 , there is a fixed probability that a Type II error will be made. This probability can be illustrated graphically, as shown in Figure 1.

Each of the two curves in Figure 1 represents the distribution of the test statistic (e.g., the computed z value if a z test is being conducted). The curve on the left is labeled H_0 because it represents the distribution of the test statistic that would occur if the null hypothesis happened to be true. The curve on the right is labeled H_a because it represents the distribution of the test statistic that would occur if the null hypothesis happened to be false by an amount equal to the horizontal distance between the two curves.

In Figure 1, there is a vertical line that simultaneously divides each curve into two parts. The position of that line was determined by considering only (a) the H_0 curve, (b) the directional or nondirectional nature of the alternative hypothesis, and (c) the level of significance. Because the illustration in Figure 1 has been set up to depict a one-tailed test conducted with α set equal to 0.05, the vertical line is positioned such that it creates, in the H_0 curve, a tail consisting of 5% of the full H_0 distribution. This portion of the baseline to the right of the vertical line is referred to as the *region of rejection*. Once computed, the data-based test statistic is simply compared to the numerical value on the baseline under the vertical line to see if the former exceeds the latter. If so, H_0 is rejected; if not, H_0 is not rejected.

The shaded part of the H_a curve represents the probability that a Type II error will occur. This portion of the right-hand curve shows the likelihood of

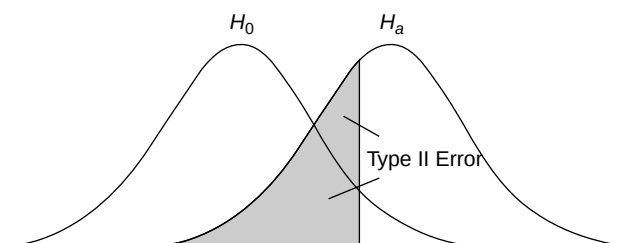


Figure 1 Type II Error Probability

not rejecting H_0 if the true state of affairs corresponds with H_a rather than H_0 . Because the shaded area of the right-hand curve constitutes 35% of the H_a curve, the probability of incorrectly failing to reject H_0 is 0.35.

Type II Error Probability and Power

In hypothesis testing, H_0 is either rejected or not rejected. If the null hypothesis is false, failing to reject it constitutes a mistake, a Type II error, while rejecting H_0 represents a correct decision. The probability that H_0 , if false, will be correctly rejected is called the statistical test's *power*.

Because the final decision in hypothesis testing is binary (either reject or do not reject H_0), one of the two events is guaranteed to occur. Therefore, the probability of not rejecting H_0 plus the probability of rejecting it is equal to 1.0. This is true regardless of whether H_0 is true or false. However, when H_0 is false, the first of these two probabilities is the probability of a Type II error, whereas the second one is the probability of a correct decision. Because the second probability is, by definition, the test's power, it follows that the probability of a Type II error + Power = 1. Rearranging terms, it also follows that Power = 1 – the probability of a Type II error. If the probability of a Type II error is represented by the word *beta* (because the terms *beta error* and *Type II error* are synonyms), it can be said that

$$\text{Power} = 1 - \text{beta}.$$

In Figure 1, the right-hand (H_a) curve is divided into two parts. The shaded part represents the probability of a Type II error (i.e., the likelihood of a beta error). The unshaded portion of the H_a curve represents the probability of correctly rejecting H_0 —the power of the statistical test. In Figure 1, power is equal to .65.

Factors That Affect the Probability of a Type II Error

When a single statistical test is run (and presuming all assumptions are met), the probability of a Type I error

is determined by one factor: the level of significance. In this same situation (one test with assumptions met), there are seven factors that jointly influence the probability of a Type II error: the numerical value in H_0 , α , the size of the sample(s), the variability in the population(s), the test used to analyze the data, the nature of H_a , and the true state of affairs in the population(s) of interest.

For any statistical test, the probability of a Type II error is lower to the extent that (a) there is a greater difference between the null hypothesis and the true state of affairs in the population(s), (b) the level of significance is lenient (i.e., high), (c) there is low variability in the population(s), and (d) large amounts of sample data are available. Also, certain statistical tests (e.g., a *t* test to compare two sample means) are less likely to produce a Type II error than are other tests (e.g., a median test). As long as they are “pointed” in the right direction, one-tailed tests are less prone to generate Type II errors than are two-tailed tests.

Because the likelihood of a Type II error is influenced jointly by many factors, any statement of Type II error risk in a given study is tied to the specific features of that particular study. Thus, the fact that Figure 1 has 35% of the H_a curve shaded is tied to a specific situation defined by the seven factors mentioned above. Make a change in any of those factors and the size of the shaded area in the H_a curve will either increase or decrease. For example, make the test two-tailed rather than one-tailed and the shaded area will increase.

Considering Type II Error Risk When Planning Studies

As indicated earlier, the most basic version of hypothesis testing involves six steps. Many researchers use that version of hypothesis testing. However, that method of research is considered by many authorities to be elementary, because no consideration is given to Type II error risk. More sophisticated studies have this kind of risk considered (and controlled) in the design phase of the investigation.

In any given study, it is possible to estimate the likelihood that a Type II error will be committed. This is done by means of a *power analysis*. In a power

analysis, a researcher indicates (or, in some cases, estimates) the seven factors mentioned in the previous section. A formula is then used to indicate the probability that a false H_0 will be rejected. That probability is the test's power. The complement of power is Type II error risk. Therefore, by doing a power analysis to see if sample sizes are large enough to produce a desired level of power, a researcher is simultaneously checking to see if the study's design is such that Type II error risk is at or below some predetermined level.

Common Misconceptions Concerning Type II Errors

There are three misconceptions that many people have about Type II errors. First, it is widely believed that Type II errors are not as “damaging” as Type I errors. (In some studies, it is far worse to draw a conclusion that is a false negative rather than a false positive.) Second, many people think that the probability of committing a Type I error is inversely related to the probability of a Type II error, with the thought being

that an increase in one brings about a decrease in the other. (Actually, it is possible to reduce both kinds of error possibilities.) Finally, it is generally thought that nonparametric tests are inherently less powerful—and thus less likely to generate Type II errors—than parametric tests. (In reality, certain nonparametric procedures, under specified conditions, have a lower Type II error risk than do their parametric counterparts.)

—Schuyler W. Huck

See also Hypothesis and Hypothesis Testing; Type I Error

Further Reading

Alexander, H. W. (1961). *Elements of mathematical statistics*. New York: Wiley.

Gravetter, F. J., & Wallnau, L. B. (2005). *Essentials of statistics for the behavioral sciences*. Belmont, CA: Wadsworth.

False negative: http://en.wikipedia.org/wiki/Type_II_error

Relationship between alpha, beta, and power: <http://www.sysurvey.com/tips/statistics/type2.htm>

Type II error: <http://www.cmh.edu/stats/definitions/typeII.htm>

U

Things could be worse. Suppose your errors were counted and published every day, like those of a baseball player.

—Author Unknown

UNBIASED ESTIMATOR

Let the data be represented by $X_1, X_2, \dots, X_n$, a collection of random variables, whose behavior can be modeled by a probability distribution F_θ , where θ represents a parameter to be estimated from the data. Here, θ may be a k -dimensional vector. For example, F_θ may be the family of normal distributions with θ representing the mean of the X_i s, or θ may be a two-dimensional vector, $\theta = (\mu, \sigma^2)$, representing the mean and variance of the X_i s. Let $\delta(X_1, \dots, X_n)$ be a function of the data. That is, $\delta(X_1, \dots, X_n)$ is an estimator.

The estimator $\delta(X_1, \dots, X_n)$ is said to be *unbiased* for $g(\theta)$ if, for all θ ,

$$E_\theta(\delta(X_1, \dots, X_n)) = g(\theta). \quad (1)$$

When Condition 1 does not hold, the *bias* of δ is defined as $b(\theta) = E_\theta(\delta(X_1, \dots, X_n)) - g(\theta)$. Gauss introduced the concept of unbiasedness to denote lack of systematic error in the estimation process.

Example 1. Let $X_1, \dots, X_n$ denote random variables with the same expectation θ . Because $E(\sum_{i=1}^n c_i X_i) = \theta$, whenever $\sum_{i=1}^n c_i = 1$, then $\sum_{i=1}^n c_i X_i$ is unbiased for θ .

Example 2 (Survey Sampling). Let a population consist of N individuals, each with annual income t_i , $i = 1, \dots, N$. To estimate the total income $T = \sum_{i=1}^N t_i$ take a random sample of size n and denote by $\{y_1, \dots, y_n\}$ the incomes of the n sampled individuals. Note that each of the $\{y_1, \dots, y_n\}$ has probability $\frac{1}{N}$ of being equal to each of the t_i , $i = 1, \dots, N$.

The following calculation shows that $\frac{N}{n} \sum_{i=1}^n Y_i$ is unbiased for T :

$$E \left\{ \frac{N}{n} \sum_{i=1}^n y_i \right\} = \frac{N}{n} \sum_{i=1}^n E(y_i) = \frac{N}{n} \sum_{i=1}^n \frac{1}{N} \sum_{j=1}^N t_j = T.$$

In most cases, the requirement of unbiasedness as defined by Condition 1 yields intuitive estimators with good properties. However, unbiased estimators may not exist, and when they do exist, sometimes they behave poorly.

Example 3. Let X be a *binomial* random variable with parameters n and $\theta =$ probability of success. It is known that there is no unbiased estimator for $g(\theta)$ unless $g(\theta)$ is a polynomial in θ of degree less than or equal to n . For example, there is no unbiased estimator for $\frac{1}{\theta}$ based on X . On the other hand, when a

geometric random variable Y with θ = probability of success is observed, Y is unbiased for $\frac{1}{\theta}$.

Example 4. Let X be a Poisson random variable with parameter θ . When estimating $(p\{X = 0\})^2 = e^{-2\theta}$, the unique unbiased estimator is $\delta(X) = (-1)^X$, a silly estimator. Lehmann considers this issue and proposes that it is due to inadequate information. For example, if, instead of a single Poisson observation, there are n independent Poisson observations, setting $T = \sum_{i=1}^n X_i$ an unbiased estimator for $e^{-2\theta}$ is $\delta(T) = (1 - \frac{2}{n})^T$, which is a reasonable estimator for $n > 2$.

A general concept of unbiasedness was proposed by Lehmann. Let $(L(\delta(X), g(\theta)))$ represent the loss incurred in estimating $g(\theta)$ by $\delta(X)$. Then $\delta(X)$ is said to be L -unbiased with respect to the loss function L if, for all $\theta^* \neq \theta$,

$$E_{\theta}(L(\delta(X), g(\theta))) \leq E_{\theta}(L(\delta(X), g(\theta^*))). \quad (2)$$

When the loss function is squared-error loss, $L(\delta(X), g(\theta)) = (\delta(X) - g(\theta))^2$, Condition 2 is equivalent to Condition 1. When the loss function is absolute error, $L(\delta(X), g(\theta)) = |\delta(X) - g(\theta)|$, Condition 2 is equivalent to *median-unbiasedness*; that is, Condition 2 is equivalent to requiring that $\delta(X)$ satisfy the condition that for all θ ,

$$P_{\theta}\{\delta(X) > g(\theta)\} = P_{\theta}\{\delta(X) < g(\theta)\}.$$

Rojo and Klebanov studied the existence of L -unbiased estimators.

—Javier Rojo

Further Reading

- Gauss, C. F. (1821). *Theoria combinationis observationum erroribus minimis obnoxiae*. Göttingen, Germany: Royal Society of Göttingen.
- Klebanov, L. B. (1976). A general definition of unbiasedness. *Theory of Probability and Its Application*, 21, 571–585.
- Lehmann, E. L. (1951). A general concept of unbiasedness. *Annals of Mathematical Statistics*, 22, 587–592.
- Lehmann, E. L. (1983). Estimation with inadequate information. *Journal of the American Statistical Association*, 78, 624–627.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer-Verlag.

Rojo, J. (1983). *On Lehmann's general concept of unbiasedness and some of its applications*. Unpublished doctoral dissertation, University of California, Berkeley.

UNIVERSAL NONVERBAL INTELLIGENCE TEST

The Universal Nonverbal Intelligence Test (UNIT) is an individually administered test designed to measure the general intelligence and cognitive abilities of children and adolescents ages 5 through 17 years who may be disadvantaged by traditional language-loaded ability tests. The UNIT is intended to provide a fair assessment for children and adolescents who have speech, language, or hearing impairments; who have different cultural or language backgrounds (e.g., English as a Second Language); and who are verbally uncommunicative due to psychiatric (e.g., elective/selective mutism), developmental (e.g., autistic spectrum), or organic reasons (e.g., traumatic brain injury).

Although its administration and response formats are entirely nonverbal (i.e., no verbal directions, no verbal items, and no verbal responses), the UNIT employs eight standardized gestures to demonstrate the nature of each task and to guide test administration. The UNIT is a comprehensive measure of intelligence and assesses a broad range of complex memory and reasoning abilities, including those lending themselves to internal processes of verbal mediation (symbolic tasks) as well as those that are less conducive to such mediation (nonsymbolic tasks). UNIT memory subtests measure complex memory, with multiple salient characteristics to be recalled (e.g., color, object, location, and sequence). The UNIT reasoning subtests measure pattern processing, problem solving, understanding of analogic relationships, and planning abilities. Although the UNIT is a highly g -saturated measure of general intelligence, with five of six subtests having average g -loadings

Table 1 UNIT Subtests, Subtest-to-Scale Assignment, and Coefficient Alpha for Scale Quotients and Full-Scale IQ

<i>SCALE/ Subtest</i>	<i>Symbolic Quotient</i>	<i>Nonsymbolic Quotient</i>	<i>Alpha</i>
Memory Quotient	<i>Symbolic Memory</i>	<i>Spatial Memory</i>	.88
	<i>Object Memory</i>		
Reasoning Quotient	<i>Analogic Reasoning</i>	<i>Cube Design Mazes</i>	.90
Alpha	.87	.91	FSIQ .93

above .70, exploratory and confirmatory factor analyses support the test's 2×2 theoretical model (i.e., Memory/Reasoning; Symbolic/Nonsymbolic). Table 1 depicts the UNIT theoretical model in its entirety, including subtest representation and total sample alpha coefficients for the UNIT scale quotients and the full-scale IQ (FSIQ). As can be seen in Table 1, the UNIT is a highly reliable and comprehensive measure of general intelligence.

For additional versatility and practicality, the UNIT combines its three memory subtests and three reasoning subtests in a flexible manner that permits the use of three possible batteries. The 15-minute UNIT Abbreviated Battery includes only the first two subtests (i.e., Symbolic Memory and Cube Design); the 30-minute Standard Battery is a four-subtest configuration that adds Spatial Memory and Analogic Reasoning to the first two subtests. The 45-minute six-subtest Extended Battery adds Object Memory and Mazes to the first four subtests. Regardless of the desired length of administration, the UNIT's theoretical model underpins each of the three batteries.

The UNIT is a standardized, norm-referenced measure. The normative data are based on a comprehensive national sample that closely matched the U.S. population on important demographic variables. The sample was composed of 2,100 children and adolescents, and included a proportional representation of students receiving services for identified exceptional needs as well as students receiving services for English as a Second Language (ESL) and bilingual education. The UNIT Examiner's

Manual dedicates an entire chapter to the topic of fairness in testing, which highlights the authors' efforts to render the instrument as fair as possible for all students. Comparative reliabilities and factor analyses by age, race/ethnicity, and gender are reported in the manual, as well as a large number of matched-sample mean score comparisons between various groups.

—Bruce A. Bracken

Further Reading

- Bracken, B. A., & Naglieri, J. A. (2003). Assessing diverse populations with nonverbal tests of general intelligence. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (2nd ed., pp. 243–274). New York: Guilford.
- McCallum, R. S., & Bracken, B. A. (2005). The Universal Nonverbal Intelligence Test: A multidimensional measure of intelligence. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 425–440). New York: Guilford.
- McCallum, R. S., Bracken, B. A., & Wasserman, J. (2001). *Essentials of nonverbal assessment*. New York: Wiley.

V

Men love to wonder, and that is the seed of science.

—Ralph Waldo Emerson

VALIDITY COEFFICIENT

The validity coefficient is a statistical index used to report evidence of validity for intended interpretations of test scores and defined as the magnitude of the correlation between test scores and a criterion variable (i.e., a measure representing a theoretical component of the intended meaning of the test). For example, the relationship between Scholastic Aptitude Test scores and college grade point average (GPA) as the criterion variable may demonstrate validity evidence for interpreting Scholastic Aptitude Test scores as an indication of academic success (intended meaning) in terms of achieving high grades (theoretical component).

According to the 1999 edition of the *Standards for Educational and Psychological Testing*, the term *validity* refers to the extent to which an intended interpretation of test scores is supported by evidence. Rather than discussing validity in terms of differing types (e.g., construct, content, and criterion-related validity), the 1999 standards view construct validity as the only type of validity that can be supported by evidence based on test content, response processes, internal structure, relations to other variables, and

consequences of testing. As empirically based forms of evidence, validity coefficients are determined by studies that examine relations to other variables (formerly referred to as criterion-related validity) and commonly include correlational studies.

Correlational studies of validity evidence often are discussed in terms of being concurrent or predictive in nature. Concurrent evidence is indicated by relations between test and criterion scores that are measured at approximately the same time, whereas predictive evidence reflects the relation between test scores gathered at one time and criterion scores measured at a later time. In other words, test scores that predict future behavior are said to have predictive validity, whereas test scores that reflect present behavior are said to have concurrent validity. For example, the relationship between Scholastic Aptitude Test scores of high school students at Time 1 and the same students' college GPAs at Time 2 would be an example of predictive validity evidence, whereas the relationship between the Scholastic Aptitude Test scores of high school students and the same students' high school GPAs would be an example of concurrent validity evidence. Table 1 presents hypothetical concurrent and predictive validity evidence for the Scholastic Aptitude Test.

Table 1 Hypothetical Concurrent and Predictive Validity Evidence for the SAT

Test	Criterion	Type of Validity	Validity Coefficient
SAT	GPA	Concurrent	.54
SAT	GPA	Predictive	.47
SAT	Class rank	Concurrent	.35
SAT	Class rank	Predictive	.32
SAT	ACT	Concurrent	.58

Less commonly used approaches for gathering evidence of the relation between test scores and criteria include the following:

studies of postdictive validity (the ability to relate to a criterion that had previously occurred)

studies of incremental validity (the ability to improve on the validity of an existing test)

studies of group comparison validity (the ability to measure expected differences for multiple groups)

studies of experimental validity (the ability to measure expected change in criteria)

Computation and Interpretation

Although there are a number of statistics for the computation of a correlation, such as the Spearman rank correlation (ρ) and Kendall's rank correlation coefficient (τ), the validity coefficient typically is computed with the Pearson product-moment correlation (r). The interpretation of validity coefficient values follows that of a correlation index. Thus, coefficient values may range from -1 to $+1$. The magnitude of the *relationship* is described in terms of strength, such that values near 0 indicate very weak relations between test and criterion scores, whereas values near $+1$ or -1 indicate very strong relations. Similarly, the magnitude of the *correlation* is discussed in terms of size (e.g., large or small). Positive values indicate direct relations (e.g., as test scores increase, so do criterion scores), whereas negative correlations indicate indirect relations (e.g., as test scores decrease, criterion scores increase).

The validity coefficient often is interpreted in terms of the square of its value (r^2). Referred to as the

coefficient of determination, the squared value represents the percentage of variance the test and criterion scores have in common. In other words, the square of the validity coefficient can be interpreted as the percentage of the variance in criterion scores that can be accounted for by test scores. For example, a validity coefficient value of .50 would indicate that 25% of the variance of the criterion (e.g., college GPA) could be accounted for by the test scores (e.g., Scholastic Aptitude Test).

In addition, the interpretation of coefficient values is often discussed with regard to statistical significance, which indicates that a true relationship between test scores and criterion scores is statistically probable. Statistical significance is a function of sample size (n), so significance can be achieved for relatively low coefficient values when such values are based on a large sample size. The statistical significance for a particular coefficient value can be determined by looking at a table of critical r values (i.e., values indicating the point when the null hypothesis should be rejected) or by computing the t statistic and looking at a table of critical t values.

The distinction between the terms *statistical significance*, *strength of the relationship*, and *meaningfulness* warrants additional discussion. Recall that a large coefficient value indicates a strong relation and that a significant value indicates that a true relationship is probable. However, a large correlation may not be significant, and a significant relationship may not be a strong one. In addition, because the validity coefficient is based on the correlation between test scores and criterion scores, a significant or strong relationship does not necessarily indicate what the test actually measures. In other words, a correlation index does not necessarily indicate a meaningful relation exists between test scores and criteria. For example, a correlation may be found between Scholastic Aptitude Test scores and shoe size, but clearly the Scholastic Aptitude Test is not a meaningful measure of shoe size. A meaningful criterion is one that is based on a sound theoretical rationale.

Last, when interpreting validity coefficient values, one should be aware that the formulas for computing correlations (e.g., Pearson's r) are based on the assumption that a linear relationship exists between

test and criterion scores. If in actuality a nonlinear (i.e., curvilinear) relationship exists, then coefficient values will be less accurate. Graphing data points in a scatter plot helps identify the linearity of a relationship.

Evaluating Validity With the Coefficient

Validity coefficient values may be evidence for the criterion validity of test score interpretation, but the evaluation of a single validity coefficient value cannot constitute an adequate validation process. Consequently, a gold standard or benchmark coefficient value that one could use to determine whether a particular interpretation of test scores is valid does not exist. Generally, however, coefficient values do not exceed .60, and values wide of .00 are commonly interpreted as indicating that test scores are superior to guesswork. To deem an interpretation to have adequate evidence of validity, one must consider other forms of information, such as the intended construct to be measured (e.g., academic achievement), the intended population of test takers (e.g., age), and the intended use of the test (e.g., college admissions). For example, the assertion that there is evidence for the validity of Scholastic Aptitude Test scores for the purpose of predicting the academic success of high school students would need to be based on an extensive examination of test scores in relation to the construct of academic success, the population of high school students, and the common use of test scores to determine college admittance.

A number of factors can influence the value of the validity coefficient. First, and perhaps most important, the meaningfulness of the criterion variable will tend to affect the coefficient value. When a criterion is not meaningful or relevant, the correlation between test scores and criterion scores likely will be weak. For instance, the validity coefficient of Scholastic Aptitude Test scores likely would be very low if the chosen criterion were a measure of shoe size. Again, meaningful criteria are based on sound theoretical rationales.

Second, the reliability of both the test scores and the criteria will influence coefficient values. This is because the validity of test scores is a function of their

reliability; that is, test scores cannot be valid if they are not reliable (although test scores may be reliable and not valid). Therefore, test scores or criteria with low levels of reliability will not be apt to achieve a high coefficient value.

Third, coefficient values also can be affected by the range of the test and criterion scores. For instance, as the range of a measure's scores increases, the corresponding coefficient values will tend to increase. The variability of criterion scores often is restricted or reduced in cases in which individuals with high test scores are selected for such things as admittance to highly competitive colleges or employment. For example, if Scholastic Aptitude Test scores of high-achieving students are collected and compared to the students' GPAs, the variability of test scores and GPAs would be relatively low (because nearly all the scores and GPAs will be very high), which would engender a low validity coefficient value. This type of situation is referred to as a *restriction of range*.

If restriction of range or a low level of reliability has occurred, correction formulas can be used to determine "corrected" validity coefficient values. An appropriate use of correction formulas would be to determine the degree that the validity coefficient would increase if the reliability for the test or criterion were to increase a given amount. This information would be helpful to test developers who may be considering a revision to an existing measure. An inappropriate use would be to evaluate the validity of test scores with corrected coefficient values. The following formula can be used to compute corrected coefficient values based on a change in the reliability of test scores or criteria:

$$r_c = r_o \left(\frac{\sqrt{\alpha_Y'} \sqrt{\alpha_X'}}{\sqrt{\alpha_Y} \sqrt{\alpha_X}} \right),$$

where

r_c is the corrected correlation value;

r_o is the observed correlation, .50;

α_Y' is the improved reliability coefficient for the criterion, .85;

α'_x is the improved reliability coefficient for the test, .89;

α_y is the measured reliability coefficient for the criterion, .75; and

α_x is the measured reliability coefficient for the test, .74.

After the values have been entered, the equation appears like this:

$$r_c = .5 \left(\frac{\sqrt{.85}\sqrt{.89}}{\sqrt{.75}\sqrt{.74}} \right) = .58.$$

Valid test scores may or may not be of value to test users; therefore, the validity coefficient alone also cannot indicate the overall value or goodness of a test. To make a judgment concerning the value of a particular test, one must examine the relevant types of validity and reliability; the relative value of a test also should be examined in terms of both the utility and cost of the test and the utilities and costs of preexisting tests. For instance, if a new measure were more costly than an existing measure but only slightly improved in terms of validity coefficient values, the new test would tend not to be of value to test users. However, if the measures were being used in high-stakes situations, the improved test might be of value.

Last, it is the responsibility of publishers and test authors to provide adequate evidence of validity and reliability for test scores. Informed decisions with regard to appropriate test selection, usage, and interpretation of test scores require test users to be familiar with the psychometric data. In their validation processes, test users can find validity coefficient values and other validity and reliability evidence in test manuals and also in studies published in peer-reviewed journals. A comprehensive validation process considers the corpus of evidence supportive and contradictory of intended, as well as alternative, interpretations of test scores.

—Shawn T. Bubany

See also Construct Validity; Content Validity; Face Validity

Further Reading

- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Haertel, E. H. (1999). Validity arguments for high-stake testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5–9.
- Silverlake, A. C. (1999). *Comprehending test manuals: A guide and workbook*. Los Angeles: Pyrezak.

Critical *r* values table: www.ecowin.org/aulas/resources/stats/correlation.htm

VALIDITY THEORY

The concept of *validity* is one of the most influential concepts in science because considerations about its nature and scope influence everything from the design to the implementation and application of scientific research. Validity is not an abstract property of any observable, unobservable, or conceptual phenomenon, such as a measurement instrument, a personality trait, or a study design. Rather, validity is a characteristic of the inferences that are drawn about phenomena by human agents and the actions that result from these inferences. Specifically, validity is always a matter of degree and not absolutes. This stems partly from the fact that validity is not an observable characteristic of inferences and actions but something that has to be inferred also.

The evaluation of the degree to which inferences are valid and resulting actions are justifiable is, therefore, necessarily embedded in a social discourse whose participants typically bring to the table diverse frameworks, assumptions, beliefs, and values about what constitutes credible evidence. Specifically, modern frameworks for validity typically list both rational and empirical pieces of evidence as necessary, but in each individual context, what these pieces should look like is open to debate. Put differently, a coherent statement about the validity of inferences and actions requires negotiation as well as consensus and places

multiple responsibilities on the stakeholders who develop such a statement.

Negotiating Validity

A metaphor may illustrate complications that can arise in a discourse about validity. If an educational assessment is viewed as the construction of a house, inferences are markers of the utility of the house. In this sense, an evaluation of the validity of inferences can be viewed as an evaluation of the degree to which the house provides structural support for the purposes that are envisioned for it. Obviously, the parties who are envisioning a certain use of the house are not necessarily the same as the designers or builders of the house, and so discrepancies can arise easily. Of course, other reasons for a mismatch are possible and could stem from a miscommunication between the designers of the house and the users of the house or from a faulty implementation of the design plans for the house. In a sense, the search for inferences that can be supported can be viewed as the search for how a house can be transformed into a home.

In general, the stakeholders in an assessment can be coarsely viewed as belonging to four complementary groups. First, there are the test developers, who create a research program, a framework, or an instrument under multiple considerations, such as theoretical adequacy and feasibility for practical implementation. Second are the examinees, whose needs in the process are typically more practical and may differ quite substantially from those of the other stakeholders involved. Third are the test users, or the decision makers who utilize the scores and diagnostic information from the assessment to make decisions about the examinees; only rarely are the examinees the only decision makers involved. Fourth are the larger scientific and nonscientific communities to which the results of an assessment program are to be communicated and whose needs are a *mélange* of those of the test developers, the test users, and the examinees. Therefore, determining the degree to which inferences and actions are justifiable is situated in the communicative space among these different stakeholders.

Not surprisingly, examples of problems in determining the validity of inferences abound. For example, the inferences that test users may want to draw from a certain assessment administered to a certain population may be more commensurate with an alternative assessment for a slightly different population. However, that is not a faulty characteristic of the assessment itself. Rather, it highlights the difference between the agents who make inferences and the agents who provide a foundation for a certain set of inferences, of which the desired inferences may not be a member.

Historical Developments of Validity Theories

Until well into the 1970s, validity theory presented itself as the coexistent, though largely unrelated, trinity of *criterion*-based, *content*-based, and *construct*-based conceptions of validity. According to the criterion-based approach, the validity of an assessment could be evaluated in terms of the accuracy with which a test score could predict or estimate the value of a defined criterion measure, usually an observable performance measure. The criterion-based model, notably introduced by Edward L. Thorndike at the beginning of the 20th century, owed much of its lingering popularity to an undisputable usefulness in many applied contexts that involve selection decisions or prognostic judgments, such as hiring and placement decisions in the workplace or medical and forensic prognoses. Depending on whether the criterion is assessed at the same time as the test or at a subsequent time, one can distinguish between *concurrent* and *predictive* validity, respectively. Though a number of sophisticated analytical and statistical techniques have been developed to evaluate the criterion validity of test scores, the standard methods applied were simple regression and correlation analyses. The resulting coefficient was labeled *validity coefficient*. Occasionally, these procedures were supplemented by the *known-groups method*. This approach bases validity statements on a comparison of mean test scores between groups with hypothesized extreme values (e.g., devout churchgoers and members of sex-chat

forums on the Internet on a newly developed sexual permissiveness scale).

The content-based model of validity comes into play when a well-defined and undisputed criterion measure is not readily available, especially when the prediction is targeted at a broader and multifaceted criterion (e.g., achievement in a content area like mathematics). An argument for content validity is usually established through a panel of experts, who evaluate the test content in terms of (a) relevance and (b) representativeness for the content area under scrutiny. Not surprisingly, the vagueness and subjectivity of the evaluation process has led many psychometricians to discount the content-based model as satisfying *face-validity* requirements at best. However, modern proponents of the content-based model have applied a wealth of sophisticated quantitative procedures to ensure and evaluate interrater agreement, thereby trying to lend credibility to otherwise qualitative and judgment-based validity evidence.

Shortcomings of the criterion-based and the content-based models of validity incited the American Psychological Association to set forth technical recommendations for justifying interpretations of psychological tests. As a result of this endeavor, the term *construct validity* was coined and later elaborated by Lee J. Cronbach and Paul Meehl. In the beginning, they tied their validity theory closely to a more general and abstract nomological network, which was described in 1952 by Carl G. Hempel in his classic essay *Fundamentals of Concept Formation in Empirical Science*. Metaphorically and graphically, the constructs are represented by knots, and the threads connecting these knots represent the definitions and hypotheses included in the theory. The whole system, figuratively speaking, “floats” above the plane of observation and is connected to it by “strings,” or rules of interpretation. The complex system of theoretical definitions can be used to formulate theoretical hypotheses, which can, in turn, be used to formulate empirical hypotheses about relationships among observable variables. In this framework, validity is not a characteristic of a construct or its observed counterpart but of the interpretation of defined logical relations of a causal nature that function

to semantically circumscribe a theoretical network of constructs and construct relations.

An obvious epistemological problem arises, however, when the observed relationships are inconsistent with theory, which is exacerbated by the dearth of developed formal theories in many psychological and social science domains. This lack of strong theory led Cronbach to coin the phrases “weak program” and “strong program” of construct validity. He cautions that, without solid theory (i.e., with only a weak program of construct validity), every correlation of the construct under development with any other observed attribute or variable could be accepted as validity evidence. Consequently, in the absence of any coordinated argument, validation research would then resemble more an empirical shotgun procedure than a scientific program.

Such problems notwithstanding, by the 1980s, the notion of construct validity became accepted as the basis for a new framework of validity assessment that is characterized by its unifying nature. The unifying aspect stems primarily from the acknowledgment that interpretive elements like assumptions and value judgments are pervasive when measuring psychological entities and, thus, are unavoidable in any discourse about any aspect of validity. As Samuel Messick, the most prominent proponent of a unified theory of validity, has framed it, “The validation process is scientific as well as rhetorical and requires both evidence and argument.”

The most controversial aspect of the unified concept of validity as developed by Messick pertains to the role of consequences in the validation process. In this view, a validity argument must specifically consider and evaluate the social consequences of test interpretation and test use, which are describable only on the basis of social values. Importantly, his notion of social consequences does not refer merely to test misuse but, specifically, to the unanticipated consequences of legitimate test score interpretation and use. A number of critics reject his idea that evidential and consequential aspects of construct validity cannot be separated, but despite this debate and recent clarifications on the meaning of the consequential aspects, the question of value justification within a unified validity approach persists.

Philosophical Challenges

The unification of validity theory under a constructivist paradigm has challenged the prevailing implicit and explicit philosophical realism that many applied social scientists had hitherto followed in their practical measurement endeavors. In philosophical realism, a test's task was to accurately measure an existing entity and not to question whether such an entity existed in the first place (an ontological question) or whether it could be assessed at all (an epistemological question). In the constructivist view, it is not a test that is validated but its interpretation (i.e., the inferences that are drawn from a test score). Therefore, it is insufficient to operationalize validity through a single validity coefficient. Rather, validation takes the form of an open-ended argument that evaluates the overall plausibility of the proposed test score interpretations from multiple facets. Currently, the strengthening of cognitive psychology principles in construct validation as described by Susan Embretson and Joanna Gorin, for example, appears to be one of the most promising avenues for developing validity theory toward a more substantive theory that can truly blend theoretical models with empirical observations. Models with genesis in cognitive psychology enable one to disentangle and understand the processes that respondents engage in when they react to test items and to highlight the test instrument as an intervention that can be used to search for causal explanations, an argument that was developed recently in detail by Borsboom, Mellenbergh, and van Heerden.

Perspectives for the Future

To comprehensively develop a unified theory of validity in the social sciences, a lot more must be accomplished besides a synthesis of the evidential and consequential bases of test interpretation and use. In particular, a truly unified theory of validity would be one that crosses methodological boundaries and builds on the foundations that exist in other disciplines and subdisciplines. Most prominently, consider the *threats-to-validity* approach for generalized causal inferences from experimental and

quasi-experimental designs, the closely related *validity generalization* approach by virtue of meta-analytical techniques, and the long tradition of validity concepts in qualitative research. In the end, it may be best to acknowledge that validity itself is a complex construct that also needs to be validated every once in a while.

—*André A. Rupp and Hans Anand Pant*

See also Construct Validity; Content Validity; Face Validity; Validity Coefficient

Further Reading

- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061.
- Cureton, E. E., Cronbach, L. J., Meehl, P. E., Ebel, R. L., Ward, A. W., & Stoker, H. W. (1996). Validity. In A. W. Ward, H. W. Stoker, & M. Murray-Ward (Eds.), *Educational measurement: Origins, theories, and explanations: Vol. 1. Basic concepts and theories*. Lanham, MD: University Press of America.
- Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343.
- Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research*, 45, 7–34.
- Murphy, K. R. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Erlbaum.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, 11(4), 522–537.

VARIABLE

A variable represents a class of outcomes that can take on more than one value. For example, car make is a variable that can take on the values of Pontiac,

Volvo, or Chevrolet, among others. Other examples of variables are height (expressed as short or tall or 5 feet or 6 feet, etc.), income (expressed as more than \$60,000 or less than \$60,000, for example), age at menarche, number of badges earned, time in rank, speed for the 100-yard dash, and favorite type of food. All these characteristics can take on any one of several values.

Table 1 lists types of variables, definitions, and alternative terms for the variables.

The two most important variables are dependent and independent variables.

A dependent variable represents the measure that reflects the outcomes of a research study. For example, if the difference between two groups of adults on how well they can remember a set of 10 single digits after a 5-hour period is measured, the number of digits remembered is the dependent variable.

A dependent variable is the outcome measure that depends on the experimental treatment or on what the researcher changes or manipulates.

An independent variable represents the treatments or conditions that the researcher controls, either directly or indirectly, to test their effects on a particular outcome. An independent variable is also known as a treatment variable; it is in this context that the term *treatment* is most often used. An independent variable is manipulated in the course of an experiment to understand the effects of this manipulation on the dependent variable.

For example, the effectiveness of three different reading programs on children's reading skills may be tested. Method A includes tutoring, Method B includes tutoring and rewards, and Method C includes neither tutoring nor rewards (these kids just spend some time with the teacher). In this example, the method of reading instruction is manipulated, and it is the independent variable. The outcome, or dependent, variable could be reading scores. This experiment includes three levels of one independent variable (method of teaching) and one dependent variable (reading score).

The distinction between direct and indirect manipulation of the independent variable(s) has to do with whether the researcher actually creates the levels (such as Method A, Method B, and Method C above) or whether the levels occur naturally and cannot be manipulated directly but only tested. Examples of naturally occurring variables include differences such as gender (we cannot very well assign that trait to people) or age (we cannot make people younger or older).

The general rule to follow in experimental design is that when the researcher is manipulating anything or assigning participants to groups based on some characteristic, such as age or ethnicity or treatment, that variable is the independent variable. When researchers look at some outcome to determine whether the grouping had an effect, they are looking at the dependent variable.

Table 1 Types of Variables

Type	Definition	Other Terms You Might See
Dependent variable	A variable that is measured to see whether the treatment or manipulation of the independent variable had an effect	Outcome variable Results variable Criterion variable
Independent variable	A variable that is manipulated to examine its impact on a dependent variable	Treatment Factor Predictor variable
Control variable	A variable that is related to the dependent variable, the influence of which needs to be removed	Restricting variable
Extraneous variable	A variable that is related to the dependent variable or independent variable and that is not part of the experiment	Threatening variable
Moderator variable	A variable that is related to the dependent variable or independent variable and has an impact on the dependent variable	Interacting variable

Independent variables must take on at least two levels or values (because they are variables). For example, if a researcher were studying the effects of gender differences (the independent variable) on language development (the dependent variable), the independent variable would have two levels, male and female. Similarly, if a researcher were investigating differences in stress for people 30 to 39 years of age, 40 to 49 years, and 50 to 59 years, then the independent variable would be age, and it would have three levels.

The Relationship Between Independent and Dependent Variables

The best independent variable is one that is independent of any other variable that is being used in the same study. In this way, the independent variable can contribute the maximum amount of understanding beyond what other independent variables can offer. When variables compete to explain the effects, their potential influence is sometimes called confounding.

The best dependent variable is one that is sensitive to changes in the different levels of the independent variable.

—Neil J. Salkind

See also Dependent Variable; Descriptive Research; Independent Variable

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics* (2nd ed.). Thousand Oaks, CA: Sage.

HyperStat Online tutorial and steps for the computation of the variance: <http://davidmlane.com/hyperstat/A29697.html>

VARIABLE DELETION

The law of parsimony (see “Ockham’s Razor”) states that the fewer variables used to explain a situation, the more probability that the explanation will be closer to

reality, which confers more replicability and more generalizability on the explanation. Reducing the number of variables lessens Type II error probability because degrees of freedom are also lessened. The goal is to estimate a given amount of variance with the smallest variable set possible; thus bigger is not better when one is using canonical correlation analysis (CCA), making variable deletion a necessity.

The following example uses CCA and variable deletion strategies (DSs) to illustrate how to arrive at the most parsimonious variable set for predicting how well students’ attitudes toward mathematics correlate with their geometric and spatial visualization abilities. Students were administered three tests: (a) a math attitude survey containing six subscales, usefulness (U), intrinsic value (I), worry (W), confidence (C), perceptions (P), and attitude (A) toward success, serving as the six predictor variables; (b) a spatial relationship test assessing spatial sense; and (c) a geometry content knowledge test ranging from Level 0 to Level 2. The two math tests and Level 0 of the geometry test served as the three criteria variables (space relationship test, or space rel; variable Level 0, or Lev 0; and geometry content knowledge score, or GCK sum). The command syntax for running the MANOVA analysis in SPSS was as follows:

MANOVA

```
spacerelev0 gcksum with U I W C P A /
print=signif (multiv eigen dimenr)
discrim (stan estim cor) alpha (.999))/design.
```

The results of the analysis are shown in Table 1.

To make the process of completing the table as clear as possible, the “func” (canonical function coefficient), the “ r_s ” (canonical structure coefficient), and the R_c^2 (squared canonical correlation coefficient) for each function were gotten directly from the SPSS printout. The r_s^2 (squared canonical structure coefficient) was figured by squaring the canonical structure coefficients for each variable and putting them in percentage format. The h^2 (communality coefficient) for each variable was obtained by summing all the r_s^2 s. The adequacy coefficient, an average of all the squared structure coefficients for the variables in one set with respect to one function, was calculated by

Table 1 Initial Solution With Canonical Commuality Coefficients (h^2)**DS 1**

Statistic	Function 1			Function 2			Function 3			h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	0.5	-0.845	71.40%	0.556	0.162	2.62%	0.956	0.509	25.91%	99.94%
Lev0	-0.179	-0.604	36.48%	1.008	0.510	26.01%	-0.617	-0.613	37.58%	100.07%
gcksum	-0.521	-0.901	81.18%	-1.197	-0.331	10.96%	-0.843	-0.279	7.78%	99.92%
Adequacy			63.02%			13.20%			23.76%	
Rd			16.13%			0.49%			0.50%	
R_c^2			25.60%			3.70%			2.10%	
Rd			6.86%			0.68%			0.23%	
Adequacy			26.80%			18.35%			10.71%	
U	0.157	0.581	33.76%	0.153	-0.076	0.58%	-0.565	-0.463	21.44%	55.77%
I	-0.096	0.426	18.15%	-0.579	-0.63	39.69%	-0.862	-0.571	32.60%	90.44%
W	-0.187	-0.081	0.66%	-0.829	-0.805	64.80%	0.531	0.292	8.53%	73.99%
C	0.932	0.972	94.48%	-0.023	-0.207	4.28%	0.787	0.083	0.69%	99.45%
P	0.046	0.244	5.95%	0.2	-0.061	0.37%	0.145	0.033	0.11%	6.43%
A	0.061	0.279	7.78%	0.229	-0.061	0.37%	-0.222	-0.096	0.92%	9.08%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical commuality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; A = attitude toward success subscale.

adding all the structure coefficients in the criterion set, dividing by the number of variables in the set, and placing the result in percentage format. The adequacy coefficient for the predictor set was determined by the same method. The redundancy coefficient (the measure of the proportion of the variance of the criterion variable set predictable from the predictor variable set) was calculated by multiplying the adequacy coefficient by the R_c^2 for each function.

After examining the full CCA, the law of parsimony needed to be invoked through variable deletion. In order to make this deletion process as understandable as possible, three different strategies will be examined.

During the deletion process, three coefficients will be looked at:

- r_s^2 : how much variance a variable linearly shares with a canonical variable
- h^2 : sum of all r_s^2 s; how much of the variance in a given observed variable is reproduced by the complete canonical solution
- R_c^2 : how much each function is contributing to the overall canonical solution

DS 1 looks at the h^2 s only and involves the following steps:

1. Look at all the h^2 s.
2. Find the lowest h^2 and delete it.
3. Check the change to the R_c^2 for each function.
4. If there is little change to R_c^2 , find the next lowest h^2 .
5. Delete that variable, and repeat the process until the R_c^2 change is too big.

Limitations of DS 1 are that contributions are not evaluated until after the variable is dropped, which results in retention of a large h^2 that happened only on the last canonical function and had a small R_c^2 effect size, as shown in Tables 2 and 3.

DS 2 looks at the contribution of each function to the total canonical solution. It includes the following steps:

1. Look at the R_c^2 for each function.
2. Omit the function with the smallest R_c^2 . See Table 4.
3. Compute the subset of h^2 s.

Table 2 Solution With h^2 s With Perceptions Deleted**DS 1, Iteration 1**

Statistic	Function 1			Function 2			Function 3			h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	-0.503	-0.846	71.57%	0.522	0.142	2.02%	0.974	0.513	26.32%	99.90%
Lev0	-0.181	-0.605	36.60%	1.028	0.528	27.88%	-0.583	-0.596	35.52%	100.00%
gcksum	-0.516	-0.9	81.00%	-1.181	-0.324	10.50%	-0.524	-0.292	8.53%	100.02%
Adequacy			63.06%			13.46%			23.45%	
Rd			16.14%			0.26%			0.94%	
R_c^2			25.60%			1.90%			4.00%	
Rd			6.62%			0.35%			0.44%	
A			25.85%			18.60%			11.04%	
U	0.167	0.581	33.76%	0.211	-0.061	0.37%	-0.53	-0.467	21.81%	55.94%
I	-0.093	0.427	18.23%	-0.56	-0.622	38.69%	-0.891	-0.603	36.36%	93.28%
W	-0.177	-0.08	0.64%	-0.817	-0.825	68.06%	0.525	0.255	6.50%	75.21%
C	0.934	0.973	94.67%	-0.03	-0.204	4.16%	0.802	0.079	0.62%	99.46%
P	0	0	0.00%	0	0	0.00%	0	0	0.00%	0.00%
A	0.072	0.279	7.78%	0.286	-0.057	0.32%	-0.176	-0.098	0.96%	9.07%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical communality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; A = attitude toward success subscale.

Table 3 Solution With h^2 s With P and A Deleted**DS 1, Iteration 2**

Statistic	Function 1			Function 2			Function 3			h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	-0.504	-0.846	71.57%	0.583	0.171	2.92%	0.938	0.505	25.50%	100.00%
Lev0	-0.190	-0.610	37.21%	0.984	0.482	23.23%	-0.651	-0.628	39.44%	99.88%
gcksum	-0.509	-0.898	80.64%	-1.218	-0.349	12.18%	-0.441	-0.266	7.08%	99.90%
Adequacy			63.14%			12.78%			24.01%	
Rd			16.10%			0.43%			0.43%	
R_c^2			25.50%			3.40%			1.80%	
Rd			6.29%			0.67%			0.20%	
Adequacy			24.67%			19.78%			10.97%	
U	0.175	0.582	33.87%	0.229	-0.075	0.56%	-0.584	-0.475	22.56%	57.00%
I	-0.093	0.43	18.49%	-0.629	-0.664	44.09%	-0.845	-0.557	31.02%	93.60%
W	-0.153	-0.078	0.61%	-0.732	-0.840	70.56%	0.549	0.339	11.49%	82.66%
C	0.950	0.975	95.06%	0.082	-0.187	3.50%	0.764	0.087	0.76%	99.32%
P	0	0	0.00%	0	0	0.00%	0	0	0.00%	0.00%
A	0	0	0.00%	0	0	0.00%	0	0	0.00%	0.00%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical communality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; A = attitude toward success subscale.

Table 4 Canonical Solution After Dropping Function 3 With Subset h^2 s**DS 2, Iteration 1**

Statistic	Function 1			Function 2			h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	−0.5	−0.845	71.40%	0.556	0.162	2.62%	74.03%
Lev0	−0.179	−0.604	36.48%	1.008	0.510	26.01%	62.49%
gcksum	−0.521	−0.901	81.18%	−1.197	−0.331	10.96%	92.14%
Adequacy			63.02%			13.20%	
Rd			16.13%			0.49%	
R_c^2			25.60%			3.70%	
Rd			6.86%			0.68%	
Adequacy			26.80%			18.35%	
U	0.157	0.581	33.76%	0.153	−0.076	0.58%	34.33%
I	−0.096	0.426	18.15%	−0.579	−0.63	39.69%	57.84%
W	−0.187	−0.081	0.66%	−0.829	−0.805	64.80%	65.46%
C	0.932	0.972	94.48%	−0.023	−0.207	4.28%	98.76%
P	0.046	0.244	5.95%	0.2	−0.061	0.37%	6.33%
A	0.061	0.279	7.78%	0.229	−0.061	0.37%	8.16%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical communality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; A = attitude toward success subscale.

- Now find the variable that has the lowest h^2 ; drop it from the original solution.
- Repeat the preceding steps until the remaining variables are reasonably close in their subset h^2 values. This will be a matter of researcher judgment.

Limitations of DS 2 are that it ignores functions with small R_c^2 values and the variations as to where h^2 values come from. P, A, and U are deleted in Table 5.

DS 2 considers weighted h^2 s, looking at the variables' contribution to the complete canonical solution. It includes the following steps:

- Multiply R_c^2 times r_s^2 and add the products together for each row.
- Drop the lowest weighted h^2 .
- Look at the change in the R_c^2 ; if there is little change, drop the next-lowest h^2 .
- Repeat, taking out as many variables as possible without compromising the R_c^2 .

This is considered the best of the DSs because the h^2 s are weighted (see Table 6).

The goal of all DSs is a more parsimonious solution. Therefore, choosing the smaller variable set when the same amount of variance can be reproduced is the objective, because “bigger is not better!” in CCA.

—Mary Margaret Capraro

See also Variable

Further Reading

- Cantrell, C. (1999). Variable deletion. In B. Thompson (Ed.), *Advances in social science methodology*, Vol. 5 (pp. 321–333). Greenwich, CT: JAI Press.
- Capraro, M. M., & Capraro, R. M. (2001). Bigger is not better: Seeking parsimony in canonical correlation analysis via variable deletion strategies. *Multiple Linear Regression Viewpoints*, 27(2), 24–33.
- Stevens, J. (1999). Canonical correlations. In J. Stevens (Ed.), *Applied multivariate statistics for the social sciences* (3rd ed.; pp. 429–449). Mahwah, NJ: Erlbaum.
- Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretations*. Beverly Hills, CA: Sage.

Table 5 Solution With h^2 s With P, A, and U Deleted**DS 2, Iteration 4**

Statistic	Function 1			Function 2			h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	-0.491	-0.837	70.06%	0.692	0.225	5.06%	75.12%
Lev0	-0.216	-0.629	39.56%	0.892	0.393	15.44%	55.01%
gcksum	-0.503	-0.9	81.00%	-1.268	-0.389	15.13%	96.13%
Adequacy			63.54%			11.88%	
Rd			16.20%			0.40%	
R_c^2			25.50%			3.40%	
Rd			4.96%			0.68%	
Adequacy			19.46%			19.90%	
U	0	0	0.00%	0	0	0.00%	0.00%
I	-0.065	0.434	18.84%	-0.682	-0.709	50.27%	69.10%
W	-0.139	-0.073	0.53%	-0.68	-0.817	66.75%	67.28%
C	1.031	0.987	97.42%	0.249	-0.154	2.37%	99.79%
P	0	0	0.00%	0	0	0.00%	0.00%
A	0	0	0.00%	0	0	0.00%	0.00%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical communality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; A = attitude toward success subscale.

Table 6 Initial Solution With h^2 s**DS 3**

Statistic	Function 1			Function 2			Function 3			Weighted h^2
	Func.	r_s	r_s^2	Func.	r_s	r_s^2	Func.	r_s	r_s^2	
spacerel	-0.500	-0.845	71.40%	0.556	0.162	2.62%	0.956	0.509	25.91%	18.92%
Lev0	-0.179	-0.604	36.48%	1.008	0.510	26.01%	-0.617	-0.613	37.58%	11.09%
gcksum	-0.521	-0.901	81.18%	-1.197	-0.331	10.96%	-0.843	-0.279	7.78%	21.35%
Adequacy			63.02%			13.20%			23.76%	
Rd			16.13%			0.49%			0.50%	
R_c^2			25.60%			3.70%			2.10%	
Rd			6.86%			0.68%			0.23%	
Adequacy			26.80%			18.35%			10.71%	
U	0.157	0.581	33.76%	0.153	-0.076	0.58%	-0.565	-0.463	21.44%	9.11%
I	-0.096	0.426	18.15%	-0.579	-0.63	39.69%	-0.862	-0.571	32.60%	6.80%
W	-0.187	-0.081	0.66%	-0.829	-0.805	64.80%	0.531	0.292	8.53%	2.74%
C	0.932	0.972	94.48%	-0.023	-0.207	4.28%	0.787	0.083	0.69%	24.36%
P	0.046	0.244	5.95%	0.2	-0.061	0.37%	0.145	0.033	0.11%	1.54%
A	0.061	0.279	7.78%	0.229	-0.061	0.37%	-0.222	-0.096	0.92%	2.03%

Note: spacerel = space relations portion of differential aptitude test score; Lev0 = score from geometry content knowledge test; gcksum = geometry content knowledge test sum score; Func. = canonical function coefficient; r_s = canonical structure coefficient; r_s^2 = squared canonical structure coefficient; h^2 = canonical communality coefficient; Rd = redundancy coefficient; Adequacy = adequacy coefficient; R_c^2 = squared canonical correlation coefficient; U = usefulness subscale; I = intrinsic value subscale; W = worry subscale; C = confidence subscale; P = perceptions subscale; S = attitude toward success subscale.

VARIANCE

In addition to the standard deviation, a very common measure of variability is the variance. If you know the standard deviation of a set of scores, you can easily compute the variance of that same set of scores; the variance is simply the square of the standard deviation, as shown in the following formula:

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1},$$

where

s^2 is the variance,

$\sum$ is the summation of deviations from the mean squared,

X is the data point,

$\bar{X}$ is the mean of the set of data points,

n is the sample size.

Consider the data set of 10 observations as follows:

5, 7, 5, 3, 4, 6, 5, 7, 9, 9,

which has a mean of 6. From the above formula, the value of the variance is 4.00.

Excel's VAR Function

While variance is a relatively easy descriptive statistic to compute by hand, it is much easier to use a software package such as Excel and the VAR function.

To compute the variance of a set of numbers by means of Excel, follow these steps:

1. Enter the individual scores into one column in a worksheet. The data in this example were entered into cells A1 through A10.
2. Select the cell into which you want to enter the VAR function. In this example, we are going to compute the variance in cell A11.

	A11		f_x	=VAR(A1:A10)
	A	B	C	D
1	5			
2	7			
3	5			
4	3			
5	4			
6	6			
7	5			
8	7			
9	9			
10	9			
11	4.00			
12				

Figure 1 Using the VAR Function to Compute the Variance

3. Now click on cell A11 and type the Average function as follows,

=VAR (A1:A10),

and press the Enter key.

4. As you can see in Figure 1, the variance was computed and the value returned to cell A11. Notice that in the formula bar in Figure 1, you can see the VAR function fully expressed and the value computed.

Column1	
Mean	6.000
Standard Error	0.632
Median	5.500
Mode	5.000
Standard Deviation	2.000
Sample Variance	4.000
Kurtosis	-0.738
Skewness	0.313
Range	6.000
Minimum	3.000
Maximum	9.000
Sum	60.000
Count	10.000

Figure 2 Using the Data Analysis ToolPak to Compute the Variance and Other Descriptive Statistics

Using the Excel Data Analysis ToolPak

The variance can also be computed easily with Excel's Data Analysis ToolPak, as shown in Figure 2. Here, the Data Analysis option was selected from the Tools menu, the Descriptive Statistics option was selected, and the range of data and location for output were identified. The results show a complete descriptive analysis of the data.

—Neil J. Salkind

See also Data Analysis ToolPak; Excel Spreadsheet Functions; Standard Deviation

Further Reading

Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics* (2nd ed.). Thousand Oaks, CA: Sage.

HyperStat Online tutorial and steps for the computation of the variance: <http://davidmlane.com/hyperstat/A16252.html>

VERBAL IQ

Verbal IQ is a measure of aspects of intelligence that relate to words and language and is associated with the Wechsler Intelligence Scale. When David Wechsler developed his IQ test in 1939, he divided it into two components—tasks that required mainly verbal abilities and tasks that required mainly perceptual-manipulative skills. The verbal IQ can be considered primarily a measure of acquired knowledge, verbal reasoning, and general verbal skills. It includes measures of vocabulary knowledge, verbal concept formation, arithmetic skill, auditory memory, general fund of information, and understanding of social rules and norms. Recent editions of the Wechsler scales have broken down verbal IQ into a verbal comprehension component and a working memory component. The verbal comprehension scale most closely represents what has traditionally been thought of as verbal IQ.

Another way of thinking about verbal IQ is that it is that aspect of intelligence that depends on

experience and learning, sometimes referred to as crystallized intelligence. Because of its dependence on experience, verbal IQ is highly culturally loaded. If an individual has not had life experiences similar to the typical United States citizen, her or his score may be artificially lowered. Although the research is far from definitive, there is some indication that skills measured by verbal IQ stay the same or increase with age and experience while those measured by performance IQ tend to decline. Attempts to tie verbal IQ to the left hemisphere of the brain and performance IQ to the right hemisphere have generally not been successful.

—Steve Saladin

See also Wechsler Adult Intelligence Scale

Further Reading

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.

Kaufman, A. S., & Lichtenberger, E. O. (2005). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.

Applying Ideas on Statistics and Measurement

The following abstract is adapted from Gibson, C. L., Piquero, A. R., & Tibbetts, S. G. (2001). The contribution of family adversity and verbal IQ to criminal behavior. *International Journal of Offender Therapy and Comparative Criminology*, 45(5), 574–592.

Several studies have reported on the risk factor prevention paradigm, an effort to identify risk factors and protective factors that increase and decrease the odds of offending. For example, some have suggested that multiplicative interactions of such factors should be explored in an attempt to understand how they are linked to offending behaviors such as offending prevalence and early onset of offending. In this research, Chris Gibson and his colleagues examine Moffitt's interactional hypothesis, which states that two specific risk factors, **verbal IQ** and family adversity, interact to increase the probability of particular types of criminal behavior. Logistic regression analyses using data from the Philadelphia portion

of the Collaborative Perinatal Project of 987 African American youth indicate that the combined effect of verbal IQ and family adversity did not significantly increase the odds of becoming an offender, whereas the combined effect of low verbal IQ scores at age 7 and family adversity significantly increased the odds of early onset of offending.

VINELAND ADAPTIVE BEHAVIOR SCALES

The Vineland Adaptive Behavior Scales (VABS), published by American Guidance Service (<http://ags.pearsonassessments.com>), assess the personal and social sufficiency of individuals from birth to 90 years of age. They are used for diagnostic evaluations, program planning, and research investigations. The second edition of the Vineland Adaptive Behavior Scales (VABS-II) is currently in press and will be available by early 2006. The VABS are commonly used in conjunction with assessments of intellectual functioning to differentiate between individuals' intellectual ability and their everyday functioning. Intellectual ability scores that are within the intellectually handicapped range may not accurately describe individuals' abilities to fend for themselves in daily life. Therefore, an assessment of adaptive functioning is necessary to measure everyday living skills and design appropriate supportive interventions.

There are four versions of the VABS-II, all administered to a person who is familiar with the behavior of the individual of interest. The Survey Interview is a semistructured interview (requiring up to 60 minutes for administration) that provides a general assessment of adaptive functioning. The Expanded Interview (which takes up to 90 minutes to administer) offers a more comprehensive assessment and a systematic basis for preparing individual educational, habilitative, or treatment programs. The Parent/Caregiver Rating Form covers the same content as the Survey

Interview but uses a rating scale format and works well when time or access is limited. The Teacher Rating Form is independently completed by a classroom teacher, provides an assessment of behavior in the classroom, takes up to 20 minutes to complete, and applies to people from the age of 3 years to 22 years. Scoring can be done manually or by computer program and provides a profile of strengths and weaknesses across the domains measured.

Behaviors measured are divided into five domains: communication, daily living skills, socialization, motor skills, and maladaptive behaviors. The motor skills domain applies only to children under the age of 5 years, and the maladaptive behaviors domain applies only to children over the age of 5 years. The first four domains each have subdomains, and each subdomain has categories. For example, the daily living skills domain has a personal subdomain, which includes the categories of eating and drinking, toileting, dressing, bathing, grooming, and health care. Items are short sentences describing behaviors (for example, "Opens mouth when spoon with food is presented" or "Makes own bed when asked"), and the interviewer seeks behaviors that are regularly performed, not those that could be performed.

In the VABS-II, updating of content reflects tasks and daily living skills that are more attuned to current societal expectations than are those of the 1984 edition. Current norms match the latest census data in the United States, and current items encompass a wider range of functioning and a wider age range than did those of the first edition.

There are considerable data to support the reliability, validity, and cross-cultural stability of the VABS. They are the most highly rated scales of adaptive functioning in the United States and have been endorsed as a measure of adaptive behavior by the World Health Organization. A particular strength of the instrument is that, due to the interview structure, valuable qualitative information may be gleaned from the open-ended questions that are used to elicit specific information.

—*Fran Vertue*

Further Reading

- Beail, N. (2003). Utility of the Vineland Adaptive Behavior Scales in diagnosis and research with adults who have mental retardation. *Mental Retardation*, 41(4), 286–289.
- De Bildt, A., Kraijer, D., Sytema, S., & Minderaa, R. (2005). The psychometric properties of the Vineland Adaptive Behavior Scales in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders*, 35, 53–62.
- Sparrow, S. S., & Cicchetti, D. V. (1989). The Vineland Adaptive Behavior Scales. In C. S. Newmark (Ed.), *Major psychological assessment instruments*, Vol. 2 (pp. 199–231). Needham Heights, MA: Allyn & Bacon.

Vineland Adaptive Behavior Scales product information:
<http://ags.pearsonassessments.com>

VINELAND SOCIAL MATURITY SCALE

The Vineland Social Maturity Scale (American Guidance Service, Inc., <http://ags.pearsonassessments.com>) is an assessment scale of personal and social skills pertaining to individuals from birth to 18 years, 11 months (including low-functioning adults). This evaluation measures four domains of adaptability, including communication, daily living, socialization, and motor skills, through semistructured interviews administered to the primary caregiver. This scale is one of the most common measures of adaptive behavior and is widely used to assess individuals with mental retardation or individuals who have difficulty performing in testing situations. The Vineland is used to identify not only individuals with mental retardation but also those with developmental delays, autism spectrum disorders, and other impairments, such as attention deficit/hyperactivity disorder, and also helps in the development of educational and treatment plans.

The Vineland Social Maturity Scale was originally developed and published in 1935 by Edgar A. Doll, who was one of the first people to define mental retardation as being characterized by a limitation in adaptive skills. The scale was revised in 1984 by Sara Sparrow, Domenic Cicchetti, and David Balla and renamed the Vineland Adaptive

Behavior Scales (VABS). The most current revision is the Vineland Adaptive Behavior Scales II (VABS-II). This latest revision includes a fifth domain that encompasses maladaptive behavior, covers an age range from birth to age 90, and has updated content regarding social expectations of tasks and living skills, as well as new norms based on current U.S. census data. The VABS is available in three formats: the expanded forms, which contain 577 items and take 60 to 90 minutes to administer; the survey form, which includes 297 items and takes 20 to 60 minutes to complete; and the classroom edition of 244 items, which is a questionnaire completed by an individual's teacher and takes approximately 20 minutes.

Items address the ability to perform tasks such as dressing or preparing meals, following rules, and building and maintaining relationships. Each of the domains contains two or three subdomains, which classify the behavior into specific categories. Communication comprises receptive, expressive, and written categories. Daily living skills include personal, domestic, and community categories. Socialization consists of interpersonal relationships, play, and leisure-time categories, and coping skills and motor skills are measured as fine and gross. The new maladaptive behavior index is an optional scale of internalizing, externalizing, and other behaviors that are assessed by questions that address inappropriate or dangerous behaviors. The individual receives a score in each of the domains and an Adaptive Behavior Composite with a mean of 100 and a standard deviation of 15. A score of 70 would suggest a diagnosis of mild mental retardation, and a score of 20 to 25 would be required for a diagnosis of profound mental retardation.

—Kirsten Wells

See also Vineland Adaptive Behavior Scales

Further Reading

- Beail, N. (2003). Utility of the Vineland Adaptive Behavior Scales in diagnosis and research with adults who have mental retardation. *Mental Retardation*, 41(4), 286–289.



One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers.

—Heinrich Hertz

WECHSLER ADULT INTELLIGENCE SCALE

The Wechsler Adult Intelligence Scale (WAIS) is a set of tests that takes 90 to 120 minutes to administer and that assesses the intellectual functioning of adolescents and adults ages 16 to 89. It is the most widely used measure of adult intelligence, and its sister measures, the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI), are the most widely used for school-age children (6–16) and preschool children (ages 2–7), respectively.

The WAIS uses deviation IQ scores such that about 50% of people fall between 90 and 110, with 100 being the mean. Truly exceptional scores are 70 and below (possible mental retardation) and 130 and above (gifted). The test is divided into verbal subtests, which are designed to measure aspects of intelligence that rely on language (word meanings, general fund of knowledge, abstract verbal reasoning, etc.), and performance subtests, which measure reasoning that does not rely on language (visual pattern recognition, novel

problem solving, attention to visual detail, etc.). The Full Scale IQ (FSIQ) score is based on 11 subtests (6 verbal and 5 performance) and is generally considered the best measure of overall intelligence, unless there are large differences in performance among the subtests. Also generated are Verbal IQ and Performance IQ scores. Large differences between these scores (and between the subtests that make them up) can suggest strengths and weaknesses and in some cases neurological impairment.

The third edition (called WAIS-III and published in 1997) saw the addition of several supplemental scales that could be administered to generate index scores that were identified via factor analysis. These include the Verbal Comprehension Index (similar to Verbal IQ), Perceptual Organization Index (similar to Performance IQ), Working Memory Index (a measure of how much information someone can hold in awareness while solving problems), and Processing Speed Index (a measure of how quickly someone can perform routine visual tasks). Despite these changes, WAIS-III maintains the same basic structure as previous versions, allowing users to apply research with all versions as an aid in interpretation. The WAIS offers a

reliable and valid measure of intellectual functioning, with a large and representative standardization sample, and is used widely in research and clinical settings.

—Steve Saladin

See also Factor Analysis; Mean; Performance IQ; Verbal IQ; Wechsler Preschool and Primary Scale of Intelligence

Further Reading

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.

Kaufman, A. S., & Lichtenberger, E. O. (2005). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.

Wechsler Adult Intelligence Scale page: http://www.psychcentral.com/psypsych/Wechsler_Adult_Intelligence_Scale

WECHSLER INDIVIDUAL ACHIEVEMENT TEST

The Wechsler Individual Achievement Test (WIAT; published by Harcourt Assessment, www.harcourtassessment.com) is a comprehensive measure of basic scholastic skills. The second edition (WIAT-II) is a general broadening and updating of the original WIAT. The WIAT-II is composed of nine subtests divided into four areas: reading, mathematics, written language, and oral language. The subtest scores yield composite scores for each area, as well as a total composite score. Each subtest is organized around grade levels, from pre-kindergarten to 16 (i.e., college students and adults); not all the tests are administered to pre-kindergarten and kindergarten children. The WIAT can be administered to individuals from ages 4 to 85 and requires between 45 minutes and 2 hours to complete. The test is suitable for individual administration only.

The reading composite score is made up of the word reading, reading comprehension, and pseudo-word decoding subtests. This area includes reading aloud from lists of words, decoding pseudowords, and answering questions on given comprehension passages. The mathematics composite score is derived

from the numerical operations and math reasoning subtests. This area includes straightforward calculations as well as word problems. The written language composite score is composed of the spelling and written expression subtests. This area includes simple spelling as well as sentence, paragraph, and essay composition. The oral language composite score contains the listening comprehension and oral expression subtests. This area tests understanding of orally presented material and the ability to generate words in a category for describing a complex pictorial scene and to give verbal directions for simple tasks. The total composite score gives a general overall assessment of achievement level. The raw scores from each subtest and the composite are converted to standard scores with a mean of 100 and standard deviation of 15. Average is defined as 90 (25th percentile) to 110 (75th percentile), and by increments of 10, three additional categories on each side are defined. Grade equivalents are also available.

The WIAT-II is useful for determining areas of strength or weakness, as well as ability-achievement discrepancies that may indicate a learning disability. To assist with discrepancy analysis, it has been linked through the normative sample to the Wechsler Adult Intelligence Scale-III, the Wechsler Intelligence Scale for Children-IV, and the Wechsler Preschool and Primary Scale of Intelligence-III. In the second edition, the listening comprehension subtest of the WIAT contains only receptive vocabulary, expressive vocabulary, and sentence comprehension. It does not test the comprehension of longer, more complex passages, as was done by the first edition, so it may overestimate listening comprehension for children with attentional or memory problems. Finally, given the increasing demands of modern curricula, caution must be used when achievement expectations as expressed by these test scores are transferred to the classroom, particularly when scores fall at the low end of the average range.

—John R. Reddon and
Vincent R. Zalcik

See also Achievement Tests; Intelligence Tests

Further Reading

Lichtenberger, E. O., & Smith, D. B. (2005). *Essentials of WIAT-II and KTEA-II assessment*. Hoboken, NJ: Wiley.

Wechsler Individual Achievement Test page: <http://www.wiat-ii.com/>

WECHSLER PRESCHOOL AND PRIMARY SCALE OF INTELLIGENCE

The third edition of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III; published by the Psychological Corporation, www.harcourtassessment.com) is an individually administered, norm-referenced clinical instrument for assessing the intelligence of children between the ages of 2 years 6 months (2:6) and 7 years 3 months (7:3). The scale provides subtest and composite scores that represent intellectual functioning in verbal and performance cognitive domains, as well as a composite score that represents a child's general intellectual ability.

The scale consists of 14 subtests, grouped into three general categories: core, supplemental, and optional. The age range has been divided into two age bands to accommodate the substantial changes in children's cognitive development between the ages 2:6 and 7:3. For children age 2:6 through 3 years 11 months (3:11), the core verbal subtests are Receptive Vocabulary and Information. For the Receptive Vocabulary subtest, children look at pictures and are instructed to "Show me the cup" and "Show me raining." The core performance subtests are Block Design and Object Assembly. For Object Assembly, a child is presented with pieces of a puzzle and instructed to fit the pieces together in 90 seconds. Picture Naming is a supplemental verbal subtest. Four composite scores are possible for this age band: Verbal IQ, Performance IQ, Full Scale IQ, and General Language Composite.

For children age 4 years through 7:3, the core verbal subtests are Information, Vocabulary, and Word Reasoning. For the Information subtest, a

child responds to a question such as "Show me your nose" and "How many days make a week?" The core performance subtests are Block Design, Matrix Reasoning, and Picture Concepts. For the Matrix Reasoning subtest, the child looks at an incomplete matrix of pictures and chooses the missing portion from 4 or 5 choices. Coding is the core Processing Speed subtest. The supplemental verbal subtests are Comprehension and Similarities, and supplemental performance subtests are Picture Completion and Object Assembly. Symbol Search is the supplemental processing-speed subtest. Receptive Vocabulary and Picture Naming are optional verbal subtests but cannot be substituted for core Verbal subtests. Five composites are possible for this age band: Verbal IQ, Performance IQ, Processing Speed Quotient, Full Scale IQ, and General Language Composite.

For children age 2:6 to 3:11, administration of the core subtests takes 30 to 35 minutes; if the supplemental subtest is administered as well, 5 to 7 minutes should be added to the testing time. For children age 4 years to 7:3, administration of the core subtests takes 40 to 50 minutes; if all subtests are administered, an additional 30 to 35 minutes will be required.

The WPPSI-III can be used to obtain a comprehensive assessment of general intellectual functioning. The scale can also be used as part of an assessment to identify intellectual giftedness, cognitive developmental delays, and mental retardation. In addition, the results can serve as a guide for placement decisions in clinical or school-related programs.

—Ann M. Weber and
Kristen M. Kalymon

Further Reading

Hamilton, W., & Burns, T. G. (2003). WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence (3rd ed.). *Applied Neuropsychology*, 10(3), 188–190.

Wechsler Preschool and Primary Scale of Intelligence page: <http://harcourtassessment.com/haiweb/Cultures/en-US/dotCom/WPPSI-III/WPPSI-III.htm>

Applying Ideas on Statistics and Measurement

The following is adapted from Farver, J. A. M., Kim, Y. K., & Lee-Shin, Y. (2000). Within cultural differences: Examining individual differences in Korean American and European American preschoolers' social pretend play. *Journal of Cross-Cultural Psychology*, 31(5), 583–602.

The assessment of intelligence has been the focus of thousands of papers and countless discussions in the area of measuring and understanding individual differences. This study by Jo Ann Farver and her colleagues examined such differences in 30 Korean American and 30 European American preschoolers' play behavior to better understand how intracultural variations in children's skills and behavioral characteristics are associated with social pretend play. Children's social behaviors and play complexity were observed and recorded, teachers rated children's social behavior, parents completed a child rearing questionnaire, and children were given the **Wechsler Preschool and Primary Scale of Intelligence** and the Multidimensional Stimulus Fluency Measure. The findings showed similar patterns predictive of pretend play for both groups. Overall, children's interactive style, positive social interaction with peers, and creativity scores significantly predicted pretend play.

WEST HAVEN-YALE MULTIDIMENSIONAL PAIN INVENTORY

The West Haven-Yale Multidimensional Pain Inventory (WHYMPI) is a 52-item, multidimensional measure of chronic pain that assesses status in the following domains: perceived pain intensity and impact of pain on various facets of a patient's life; perception of responses from significant others; participation in common activities; and activity level. Each domain is divided into 12 subscales. The first domain includes five subscales: perceived interference of pain in various aspects of functioning, support and concern of significant others, pain severity, self-control, and negative mood. The second domain includes three subscales: perceived frequency of punishing responses, perceived frequency of solicitous responses, and

perceived frequency of distracting responses from significant others. The third domain includes four subscales: participation in household chores, outdoor work, activities away from home, and social activities (these scores are added for a general activity score). Each item is answered on a 7-point scale that ranges from 0 (*never*) to 6 (*very frequently*). The WHYMPI is self-administered in approximately 10 to 20 minutes. Alternative versions offer a 61-item and a 48-item Multidimensional Pain Inventory. A substantial body of research literature is available to guide the WHYMPI's applications.

Validation of the original WHYMPI was conducted on a predominantly male population of veterans experiencing a broad array of chronic pain syndromes, with lower back pain afflicting 36% of the sample. Reliability and validity estimates have been reported for men and women and across various ages and pain conditions. The WHYMPI displays internal consistency reliability ranging from 0.70 to 0.90, with test-retest stability (generally with a 2-week interval) ranging from 0.62 to 0.91. Additional normative information for particular types and locations of pain is available at www.pain.pitt.edu/mpi/MPI_Norms.pdf. Research suggests the combination of the WHYMPI, the Beck Depression Inventory, the McGill Pain Questionnaire, and a clinical interview may provide the most useful description of a patient's pain experience with respect to the four-factor model of pain: affective distress, support from significant others, pain description, and functional capacity. The activities subscale of the WHYMPI may be compromised in predicting functional capacity because some items may be culturally related or gender related. Examiners may find it helpful to inquire whether an activity was prominent prior to the onset of pain. The WHYMPI has been validated in German, Swedish, Dutch, and English, its original format. A significant-other version of the WHYMPI has been developed to assess significant others' perception of the number of solicitous, distracting, and negative responses given by a chronic pain patient.

The instrument is available from Robert D. Kerns, one of its authors, and from the ProQolid Quality of

Life Instruments Database at www.qolid.org for a minimal fee.

—Luci A. Martin and Joseph A. Doster

Further Reading

- Burckhardt, C. S., & Jones, K. D. (2003). Adult measures of pain: The McGill Pain Questionnaire (MPQ), Rheumatoid Arthritis Pain Scale (RAPS), Short-Form McGill Pain Questionnaire (SF-MPQ), Verbal Descriptive Scale (VDS), Visual Analog Scale (VAS), and West Haven-Yale Multidisciplinary Pain Inventory (WHYMPI). *Arthritis & Rheumatism (Arthritis Care & Research)*, 49, 96–104.
- De Gagne, T. A., Mikail, S. F., & D'Eon, J. L. (1995). Confirmatory factor analysis of a 4-factor model of chronic pain evaluation. *Pain*, 60, 195–202.
- Kerns, R. D., & Rosenberg, R. (1995). Pain-relevant responses from significant others: Development of a significant-other version of the WHYMPI scales. *Pain*, 61, 245–249.
- Kerns, R. D., Turk, D. C., & Rudy, T. E. (1985). The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). *Pain*, 23, 345–356.

WILCOXON, FRANK (1892–1965)

Frank Wilcoxon was born in Glengarriffe Castle near Cork in Ireland on September 2, 1892 (his wealthy American parents had rented the castle for the occasion). However, Wilcoxon became something of a rebel, running away to sea as a teenager and spending periods as an oil-well worker and as a tree surgeon. It was not until after World War I that Wilcoxon finally went to university. He studied chemistry at Rutgers University, where he was a contemporary of the singer Paul Robeson. He obtained his MSc in 1921 and his PhD in 1924 (in physical chemistry, at Cornell University).

In 1925, Ronald Fisher had published the first edition of his classic work *Statistical Methods for Research Workers*. This was compulsory reading for Wilcoxon during his postdoctoral fellowship at the Boyce Thompson Institute for Plant Research in Yonkers, where he investigated the use of copper compounds as fungicides. Wilcoxon gained further

employment at the Institute before moving, in 1943, to become group leader of the insecticide and fungicide laboratory at American Cyanamid, where he contributed to the development of malathion.

Wilcoxon's first statistical papers appeared in 1945, and it was in the second of these that he introduced the rank-sum and signed-rank tests that nowadays bear his name. These tests, and their extensions, were collected in the 1947 pamphlet titled *Some Rapid Approximate Statistical Procedures*.

In 1960, following retirement from industry, Wilcoxon accepted a part-time appointment in the new department of statistics at Florida State University, in Tallahassee. Wilcoxon had been an enthusiastic cyclist in his younger days, and at Tallahassee his preferred method of transport was a motorcycle. He died there on November 18, 1965. His innovatory statistical work is now marked by the journal *Technometrics*, which gives an annual award in his name for the best practical applications paper.

—Graham Upton

Further Reading

- Heyde, C. C., & Seneta, C. (Eds.). (2001). *Statisticians of the centuries*. New York: Springer.
- Frank Wilcoxon biographical essay: <http://www.umass.edu/wsp/statistics/tales/wilcoxon.html>

WILCOXON MANN-WHITNEY TEST

See MANN-WHITNEY U TEST (WILCOXON RANK-SUM TEST)

WILCOXON SIGNED RANKS TEST

The Wilcoxon signed ranks test may be used as a one-sample test of location or as a test of difference in location between two dependent samples. The underlying assumptions are that the distribution is continuous and symmetric.

The Wilcoxon signed ranks procedure is primarily used as a test for location, such as the median. In the one-sample case, the one-sided test of the null hypothesis is $H_0: \phi = \text{median}$, which is tested against the alternative $H_a: \phi > \text{median}$ or $H_a: \phi < \text{median}$. The alternative hypothesis for the two-sided test is $H_a: \phi \neq \text{median}$. In the two-sample case, the Wilcoxon signed ranks test is applied to matched or paired data, such as an examination of pretest versus posttest scores. Although the Wilcoxon test may also be used as a test of symmetry, there are often more powerful procedures.

The Wilcoxon test is nonparametric. This means it preserves the Type I error rate (i.e., the false positive rate) to nominal alpha regardless of the population shape. This is a fundamental advantage over its parametric counterparts, the one-mean Student t test and the two-dependent-samples Student t test, which rely on the normality distribution assumption.

Theoretical power comparisons indicate that the asymptotic relative efficiency of the Wilcoxon signed ranks test with the Student t test is .955 for the normal distribution, but it can be as high as ∞ for nonnormal distributions. Monte Carlo simulations indicate that the spectacular power gains achieved by the independent-samples version of this test (i.e., the Wilcoxon rank-sum test; see the entry on Mann-Whitney U Test) over the Student t test are not realized although the Wilcoxon signed ranks test is often at least slightly more powerful than the t test for departures from population normality. The explanation is that in the paired-samples case, tests are conducted on the distribution of the difference scores. Subtracting a pretest score from the posttest score, for example, is a mildly normalizing procedure, producing a distribution that is less deviant than the parent population and more similar to the normal distribution.

To conduct the two-sample matched-pairs Wilcoxon signed ranks test, compute the difference (d) between each pair of scores, ignoring the signs. (This is achieved by taking the absolute value

$|d|$.) Next, denote the ranks of the positive differences with a + sign and the negative differences with a – sign. Finally, sum the ranks associated with the lesser-occurring sign.

Example

A claim is made by a test preparation corporation that its study program leads to a statistically significant increase in performance on a standardized college entrance exam. The pretest is the standardized score obtained by an examinee prior to the curriculum intervention, and the posttest score is the standardized score after participating in the study program. Suppose the test scores were those depicted in Table 1.

To test this claim, subtract the pretest score from the posttest score (d) and take the absolute value ($|d|$). Note that there are eight + ranks and two – ranks. For convenience, the test statistic T is based on the smaller number of signed ranks. Therefore, the sum of the ranks with negative signs is computed:

$$\begin{aligned} T &= 3 + 1 \\ &= 4. \end{aligned}$$

The critical values for a one-sided test with $n = 10$ for $\alpha = .05$ and $\alpha = .01$ are 10 and 5, respectively. The null hypothesis is rejected if T is less than or equal to the critical value. Thus, the null hypothesis is rejected

Table 1 Pretest and Posttest Scores

Pretest	Posttest	d	$ d $	Rank of $ d $	Sign	Rank of Lesser Sign
450	484	34	34	7	+	
480	470	–10	10	1	–	1
481	516	35	35	8	+	
490	527	37	37	10	+	
492	523	31	31	4	+	
493	529	36	36	9	+	
509	479	–30	30	3	–	3
531	564	33	33	6	+	
573	605	32	32	5	+	
611	622	11	11	2	+	

Note: d = posttest score – pretest score; $|d|$ = absolute value of d .

at the $\alpha = .05$ and the $\alpha = .01$ level, and the claim that the study program is effective in increasing college entrance test scores is supported.

The p value associated with $T = 4$ is .017. In comparison, the two-dependent-samples Student t test yields $t = -2.869$, $df = 9$, and $p = .019$.

SPSS

Enter the pretest scores in one column and the posttest scores in another column. Click Analyze | Nonparametric Tests | 2 Related Samples. Then, select the pretest and posttest variables and click the “→” button to move the pair into the “Test Pair(s) List” box. Click Wilcoxon | OK. The results are depicted below.

One-Sample Test

Suppose a claim to mitigate the conclusion that the study program is effective in raising test scores is made on the basis of the caliber of participants at the pretest stage. Could the study program be effective only with high-achieving students? Suppose further

Table 2 Ranks and Test Statistics

Ranks				
		<i>N</i>	<i>Mean Rank</i>	<i>Sum of Ranks</i>
Posttest- pretest	Negative ranks	2 ^a	2.00	4.00
	Positive ranks	8 ^b	6.38	51.00
	Ties	0 ^c		
	Total	10		

Notes: a. posttest < pretest; b. posttest > pretest; c. posttest = pretest.

Test Statistics ^a	
	<i>Posttest- Pretest</i>
<i>Z</i>	-2.395 ^b
Asymp. Sig. (two-tailed)	.017

Notes: a. Wilcoxon signed ranks test; b. based on negative ranks.

that the median score of the standardized test is 500. The claim can be tested by application of the Wilcoxon signed ranks test to the pretest scores, creating another column of 10 scores, each with the value of the median = 500. The sum of the lesser signed ranks is 26, and the associated p value is .878. Therefore, the claim is dismissed because there is no evidence that the group's initial scores are statistically significantly different from the population median.

Large Sample

The Wilcoxon signed ranks test can be evaluated with a large sample approximation. This becomes necessary when tabled critical values are not available. The formula is

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

—Shlomo S. Sawilowsky

Further Reading

- Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t -test. *Annals of Mathematical Statistics*, 27, 324–335.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111, 353–360.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.

WOODCOCK JOHNSON PSYCHOEDUCATIONAL BATTERY

Originally published in 1977, the Woodcock Johnson (WJ) was the first comprehensive psychoeducational battery with conormed measures of cognitive abilities, academic achievement, and interests. It was revised in 1989 to become the Woodcock Johnson-Revised (WJ-R), dropping the interest measure and restructuring the cognitive measure to reflect the Cattell-Horn

Gf-Gc model of intelligence. The current version, the Woodcock Johnson-III (WJ-III), is comprised of two separate but conormed batteries: the Tests of Cognitive Abilities and the Tests of Achievement.

The WJ-III Tests of Cognitive Abilities was developed using an integration of the original Cattell-Horn Gf-Gc model of intelligence and Carroll's Three Stratum model, the Cattell-Horn-Carroll (CHC) theory. This second-iteration test development centering on essentially the same model has resulted in an instrument that is one of the most theoretically sound measures of intelligence on the market today. Consisting of 20 subtests organized into standard and extended batteries, the WJ-III Tests of Cognitive Abilities generates scores on seven broad CHC clusters as well as measures of general intellectual ability (brief, standard, and extended). Each subtest was chosen to represent a different narrow ability as defined by CHC theory (e.g., induction, associative memory) while the seven broad clusters represent broad CHC factors (e.g., fluid reasoning, long-term retrieval).

—Steve Saladin

See also Gf-Gc Theory of Intelligence

Further Reading

- Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence*. New York: Wiley.
- Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ-III tests of achievement assessment*. New York: Wiley.
- Schrank, F. A., & Flanagan, D. P. (2003). *WJ-III clinical use and interpretation: Scientist-practitioner perspectives*. San Diego, CA: Academic Press.
- Schrank, F. A., Flanagan, D. P., Woodcock, R. W., & Mascolo, J. T. (2002). *Essentials of WJ-III cognitive abilities assessment*. New York: Wiley.

WOODCOCK READING MASTERY TESTS REVISED

The Woodcock Reading Mastery Tests Revised (WRMTR; published by American Guidance Service, <http://www.agsnet.com/>) are a set of scales that take

40–50 minutes to administer en toto and that assess reading achievement. The WRMTR contains tests of visual-auditory learning and letter identification, which comprise the Reading Readiness Cluster. The word identification and word attack tests combine to form the Basic Skills Cluster, and the word comprehension and passage comprehension tests make up the Reading Comprehension Cluster. The word comprehension test comprises three subtests: synonyms, antonyms, and analogies.

The WRMTR has two forms, only one of which (i.e., Form G) contains the Reading Readiness Cluster tests. The tests are intended for use with individuals from kindergarten through college-age students and with older adults (75 years and older). With the exception of the visual-auditory learning and word attack tests, starting point, basal, and ceiling rules are consistent throughout the battery. The two exceptions cause confusion for some examiners. Hand scoring can be time consuming and prone to error, but the available scoring software reduces scoring time by 70 to 80 percent. The WRMTR provides a full range of score types: age and grade equivalents, standard scores, percentile ranks, relative performance indices, and instructional ranges (i.e., easy, instructional level, and difficult). Standard errors of measurement allow the hand scorer to construct 68% confidence intervals to estimate the true score. Standard errors of differences, which would allow the user to estimate an examinee's strengths and weaknesses, are not available in the manual or the scoring software. Instead, the test author recommends profile analysis, which has inherent weaknesses and provides information that is prone to error.

New norms, which include only grades kindergarten through high school, were collected in 1995 to 1996 and were based on 1994 census data. Approximately 3,400 individuals were tested, though the exact number varies from one test or cluster to another. The sample was controlled for age, gender, race, region, socioeconomic status, community size, and special populations. Nevertheless, Native Americans, Asians, and Pacific Islanders are overrepresented, and the northeastern states are underrepresented in the WRMTR sample.

The author reports only split-half reliability coefficients. The median internal consistency reliability coefficient for the tests is .91 (range = .68 to .98). For the clusters, the median reliability coefficient is .95 (range = .87 to .98). For the full-scale score, the median reliability coefficient is .97 (range = .86 to .99). The lack of alternate-form and test-retest reliability estimates is a serious omission for this highly popular test battery.

Very little information on the validity of the WRMTR is presented in the manual; much of it is based on the original (1973) version of the battery. Presumably, independent research will clarify the validity of this set of instruments, but it remains incumbent on the author

of the assessment tool to provide evidence of the technical adequacy of his product.

—Ronald C. Eaves and
Thomas O. Williams, Jr.

See also Woodcock-Johnson Psychoeducational Battery

Further Reading

Woodcock, R. N. (1998). *Woodcock Reading Mastery Tests Revised/Normative update*. Circle Pines, MN: American Guidance Service.

Woodcock Reading Mastery Tests Revised research bibliography: <http://www.agsnet.com>

Z

Research is the process of going up alleys to see if they are blind.

—Marston Bates

Z SCORES

A score by itself does not tell much. If we are told that we have obtained a score of 85 on a beauty test, this could be very good news if most people have a score of 50 but less so if most people have a score of 100. In other words, a score is meaningful only relative to the means of the sample or the population. Another problem occurs when we want to compare scores measured with different units or on different populations. How to compare, for example a score of 85 on the beauty test with a score of 100 on an IQ test?

Scores from different distributions, such as the ones in our example, can be standardized in order to provide a way of comparing them that includes consideration of their respective distributions. This is done by transforming the scores into *z* scores, which are expressed as standardized deviations from their means. These *z* scores have a mean of 0 and a standard deviation equal to 1. *Z* scores computed from different samples with different units can be directly compared because these numbers do not express the original unit of measurement.

Definition of *z* Scores

In order to compute a *z* score, we start with an original score (called *Y*) obtained from a sample (or a population) with a mean of M_Y and a standard deviation of S_Y . We eliminate the mean by subtracting it from the score, which transforms the original score into a deviation from its mean. We eliminate the original unit of measurement by dividing the score deviation by the standard deviation. Specifically, the formula for calculating a *z* score is

$$Z = \frac{Y - M_Y}{S_Y}.$$

We say that subtracting the mean *centers* the distribution and that dividing by the standard deviation *normalizes* the distribution. The interesting properties of the *z* scores are that they have a zero mean (effect of centering) and a variance and standard deviation of 1 (effect of normalizing). It is because all distributions expressed in *z* scores have the same mean (0) and the same variance (1) that we can use *z* scores to compare observations coming from different distributions.

An Example

Applying the formula for a z score to a score of $Y = 85$ coming from a sample of mean $M_Y = 75$ and standard deviation $S_Y = 17$ gives

$$Z = \frac{Y - M_Y}{S_Y} = \frac{85 - 75}{17} = \frac{10}{17} = .59.$$

Effect of z Scores

When a distribution of numbers is transformed into z scores, the shape of the distribution is unchanged, but this shape is translated in order to be centered on the value 0, and it is scaled such that its area is now equal to 1.

As a practical guide, when a distribution is normal, more than 99% of the z scores lie between the values -3 and $+3$. Also, because of the central limit theorem, a z score with a magnitude larger than 6 is extremely unlikely to occur regardless of the shape of the original distribution.

Demonstrating That z Scores Have a Mean of 0 and a Variance of 1

In order to show that the mean of the z scores is equal to 0, it suffices to show that the sum of the z scores is equal to 0. This is shown by developing the formula for the sum of the z scores:

$$\begin{aligned} \sum Z &= \sum \frac{Y - M_Y}{S_Y} \\ &= \frac{1}{S_Y} \sum (Y - M_Y) \\ &= \frac{1}{S_Y} (\sum Y - NM_Y) \\ &= \frac{1}{S_Y} (NM_Y - NM_Y) \\ &= 0. \end{aligned}$$

In order to show that the variance of the z scores is equal to 1, it suffices to show that the sum of the squared z scores is equal to $(N - 1)$ (where N is the number of scores). This is shown by developing the formula for the sum of the squared z scores:

$$\begin{aligned} \sum Z^2 &= \sum \left(\frac{Y - M_Y}{S_Y} \right)^2 \\ &= \frac{1}{S_Y^2} \sum (Y - M_Y)^2. \end{aligned}$$

But $(N - 1)S_Y^2 = \sum (Y - M_Y)^2$, hence:

$$\begin{aligned} \sum Z^2 &= \frac{1}{S_Y^2} \times (N - 1)S_Y^2 \\ &= (N - 1). \end{aligned}$$

—Hervé Abdi

Appendix A

Ten Commandments of Data Collection

The following text is taken from Neil J. Salkind's best-selling introduction to statistics text, *Statistics for People Who (Think They) Hate Statistics*, 2nd edition (2004).

Now that you know how to analyze data, you would be well served to hear something about collecting them. The data collection process can be a long and rigorous one, even if it involves only a simple, one-page questionnaire given to a group of students, parents, patients, or voters. The data collection process may very well be the most time-consuming part of your project. But as many researchers do, this period of time is also used to think about the upcoming analysis and what it will entail.

Here they are: the ten commandments for making sure your data get collected in a way that they are usable. Unlike the original Ten Commandments, these should not be carved in stone (because they can certainly change), but if you follow them, you can avoid lots of aggravation.

Commandment 1. As you begin thinking about a research question, also begin thinking about the type of data you will have to collect to answer that question. Interview? Questionnaire? Paper and pencil? Find out how other people have done it in the past by reading the relevant journals in your area of interest and consider doing what they did.

Commandment 2. As you think about the type of data you will be collecting, think about where you will be getting the data. If you are using the library for historical data or accessing files of data that have

already been collected, such as census data (available through the U.S. Census Bureau and some online), you will have few logistical problems. But what if you want to assess the interaction between newborns and their parents? The attitude of teachers toward unionizing? The age at which people over 50 think they are old? All of these questions involve needing people to provide the answers, and finding people can be tough. Start now.

Commandment 3. Make sure that the data collection forms you use are clear and easy to use. Practice on a set of pilot data so you can make sure it is easy to go from the original scoring sheets to the data collection form.

Commandment 4. Always make a duplicate copy of the data file, and keep it in a separate location. Keep in mind that there are two types of people: those who have lost their data and those who will. Keep a copy of data collection sheets in a separate location. If you are recording your data as a computer file, such as a spreadsheet, be sure to make a backup!

Commandment 5. Do not rely on other people to collect or transfer your data unless you have personally trained them and are confident that they understand the data collection process as well as you do. It is great to have people help you, and it helps keep the morale up during those long data collection sessions.

But unless your helpers are competent beyond question, you could easily sabotage all your hard work and planning.

Commandment 6. Plan a detailed schedule of when and where you will be collecting your data. If you need to visit three schools and each of 50 children needs to be tested for a total of 10 minutes at each school, that is 25 hours of testing. That does not mean you can allot 25 hours from your schedule for this activity. What about travel from one school to another? What about the child who is in the bathroom when it is his turn, and you have to wait 10 minutes until he comes back to the classroom? What about the day you show up and Cowboy Bob is the featured guest . . . and on and on. Be prepared for anything, and allocate 25% to 50% more time in your schedule for unforeseen happenings.

Commandment 7. As soon as possible, cultivate possible sources for your subject pool. Because you already have some knowledge in your own discipline, you probably also know of people who work with the type of population you want or who might be able to help you gain access to these samples. If you are in a

university community, it is likely that there are hundreds of other people competing for the same subject sample that you need. Instead of competing, why not try a more out-of-the-way (maybe 30 minutes away) school district or social group or civic organization or hospital, where you might be able to obtain a sample with less competition?

Commandment 8. Try to follow up on subjects who missed their testing session or interview. Call them back and try to reschedule. Once you get in the habit of skipping possible participants, it becomes too easy to cut the sample down to too small a size. And you can never tell—the people who drop out might be dropping out for reasons related to what you are studying. This can mean that your final sample of people is qualitatively different from the sample you started with.

Commandment 9. Never discard the original data, such as the test booklets, interview notes, and so forth. Other researchers might want to use the same database, or you may have to return to the original materials for further information.

And Number 10? Follow the previous 9. No kidding!

Tables of Critical Values

The following tables are taken from Neil J. Salkind's best-selling introduction to statistics text, *Statistics for People Who (Think They) Hate Statistics*, 2nd edition (2004).

Table 1 Areas Beneath the Normal Curve

Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score	
z-score	Area	z-score	Area	z-score	Area	z-score	Area	z-score	Area	z-score	Area	z-score	Area	z-score	Area
0.00	0.00	0.50	19.15	1.00	34.13	1.50	43.32	2.00	47.72	2.50	49.38	3.00	49.87	3.50	49.98
0.01	0.40	0.52	19.50	1.01	34.38	1.51	43.45	2.01	47.78	2.51	49.40	3.01	49.87	3.51	49.98
0.02	0.50	0.53	19.85	1.02	34.61	1.52	43.57	2.02	47.83	2.52	49.41	3.02	49.87	3.52	49.98
0.03	1.20	0.54	20.19	1.03	34.85	1.53	43.70	2.03	47.88	2.53	49.43	3.03	49.88	3.53	49.98
0.04	1.60	0.55	20.54	1.04	35.08	1.54	43.82	2.04	47.93	2.54	49.45	3.04	49.88	3.54	49.98
0.05	1.99	0.56	20.88	1.05	35.31	1.55	43.94	2.05	47.98	2.55	49.46	3.05	49.89	3.55	49.98
0.06	2.39	0.57	21.23	1.06	35.54	1.56	44.06	2.06	48.03	2.56	49.48	3.06	49.89	3.56	49.98
0.07	2.79	0.58	21.57	1.07	35.77	1.57	44.18	2.07	48.08	2.57	49.49	3.07	49.89	3.57	49.98
0.08	3.19	0.59	21.90	1.08	35.99	1.58	44.29	2.08	48.12	2.58	49.51	3.08	49.9	3.58	49.98
0.09	3.59	0.60	22.24	1.09	36.21	1.59	44.41	2.09	48.17	2.59	49.52	3.09	49.9	3.59	49.98
0.10	3.98	0.61	22.57	1.10	36.43	1.60	44.52	2.10	48.21	2.60	49.53	3.10	49.9	3.60	49.98
0.11	4.38	0.62	22.91	1.11	36.65	1.61	44.63	2.11	48.26	2.61	49.55	3.11	49.91	3.61	49.98
0.12	4.78	0.63	23.24	1.12	36.86	1.62	44.74	2.12	48.30	2.62	49.56	3.12	49.91	3.62	49.98
0.13	5.17	0.64	23.57	1.13	37.08	1.63	44.84	2.13	48.34	2.63	49.57	3.13	49.91	3.63	49.98
0.14	5.57	0.65	23.89	1.14	37.29	1.64	44.95	2.14	48.38	2.64	49.59	3.14	49.92	3.64	49.98
0.15	5.96	0.66	24.54	1.15	37.49	1.65	45.05	2.15	48.42	2.65	49.60	3.15	49.92	3.65	49.98
0.16	6.36	0.67	24.86	1.16	37.70	1.66	45.15	2.16	48.46	2.66	49.61	3.16	49.92	3.66	49.98
0.17	6.75	0.68	25.17	1.17	37.90	1.67	45.25	2.17	48.50	2.67	49.62	3.17	49.92	3.67	49.98
0.18	7.14	0.69	25.49	1.18	38.10	1.68	45.35	2.18	48.54	2.68	49.63	3.18	49.93	3.68	49.98
0.19	7.53	0.70	25.80	1.19	38.30	1.69	45.45	2.19	48.57	2.69	49.64	3.19	49.93	3.69	49.98
0.20	7.93	0.71	26.11	1.20	38.49	1.70	45.54	2.20	48.61	2.70	49.65	3.20	49.93	3.70	49.99
0.21	8.32	0.72	26.42	1.21	38.69	1.71	45.64	2.21	48.64	2.71	49.66	3.21	49.93	3.71	49.99
0.22	8.71	0.73	26.73	1.22	38.88	1.72	45.73	2.22	48.68	2.72	49.67	3.22	49.94	3.72	49.99
0.23	9.10	0.74	27.04	1.23	39.07	1.73	45.82	2.23	48.71	2.73	49.68	3.23	49.94	3.73	49.99
0.24	9.48	0.75	27.34	1.24	39.25	1.74	45.91	2.24	48.75	2.74	49.69	3.24	49.94	3.74	49.99
0.25	0.99	0.76	27.64	1.25	39.44	1.75	45.99	2.25	45.78	2.75	49.70	3.25	49.94	3.75	49.99

Table 1 (Continued)

Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score		Area Between the Mean and the z-score	
z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score	z-score
0.26	10.26	0.77	27.94	1.26	39.62	1.76	46.08	2.26	48.81	2.76	49.71	3.26	49.94	3.76	49.99
0.27	10.64	0.78	28.23	1.27	39.80	1.77	46.16	2.27	48.84	2.77	49.72	3.27	49.94	3.77	49.99
0.28	11.03	0.79	28.52	1.28	39.97	1.78	46.25	2.28	48.87	2.78	49.73	3.28	49.94	3.78	49.99
0.29	11.41	0.80	28.81	1.29	40.15	1.79	46.33	2.29	48.90	2.79	49.74	3.29	49.94	3.79	49.99
0.30	11.79	0.81	29.10	1.30	40.32	1.80	46.41	2.30	48.93	2.80	49.74	3.30	49.95	3.80	49.99
0.31	12.17	0.82	29.39	1.31	40.49	1.81	46.49	2.31	48.96	2.81	49.75	3.31	49.95	3.81	49.99
0.32	12.55	0.83	29.67	1.32	40.66	1.82	46.56	2.32	48.98	2.82	49.76	3.32	49.95	3.82	49.99
0.33	12.93	0.84	29.95	1.33	40.82	1.83	46.64	2.33	49.01	2.83	49.77	3.33	49.95	3.83	49.99
0.34	13.31	0.85	30.23	1.34	40.99	1.84	46.71	2.34	49.04	2.84	49.77	3.34	49.95	3.84	49.99
0.35	13.68	0.86	30.51	1.35	41.15	1.85	46.78	2.35	49.06	2.85	49.78	3.35	49.96	3.85	49.99
0.36	14.06	0.87	30.78	1.36	41.31	1.86	46.86	2.36	49.09	2.86	49.79	3.36	49.96	3.86	49.99
0.37	14.43	0.88	31.06	1.37	41.47	1.87	46.93	2.37	49.11	2.87	49.79	3.37	49.96	3.87	49.99
0.38	14.80	0.89	31.33	1.38	41.62	1.88	46.99	2.38	49.13	2.88	49.80	3.38	49.96	3.88	49.99
0.39	15.17	0.90	31.59	1.39	41.77	1.89	47.06	2.39	49.16	2.89	49.81	3.39	49.96	3.89	49.99
0.40	15.54	0.91	31.86	1.40	41.92	1.90	47.13	2.40	49.18	2.90	49.81	3.40	49.97	3.90	49.99
0.41	15.91	0.92	32.12	1.41	42.07	1.91	47.19	2.41	49.20	2.91	49.82	3.41	49.97	3.91	49.99
0.42	16.28	0.93	32.38	1.42	42.22	1.92	47.26	2.42	49.22	2.92	49.82	3.42	49.97	3.92	49.99
0.43	16.64	0.94	32.64	1.43	42.36	1.93	47.32	2.43	49.25	2.93	49.83	3.43	49.97	3.93	49.99
0.44	17.00	0.95	32.89	1.44	42.51	1.94	47.38	2.44	49.27	2.94	49.84	3.44	49.97	3.94	49.99
0.45	17.36	0.96	33.15	1.45	42.65	1.95	47.44	2.45	49.29	2.95	49.84	3.45	49.98	3.95	49.99
0.46	17.72	0.97	33.40	1.46	42.79	1.96	47.50	2.46	49.31	2.96	49.85	3.46	49.98	3.96	49.99
0.47	18.08	0.98	33.65	1.47	42.92	1.97	47.56	2.47	49.32	2.97	49.85	3.47	49.98	3.97	49.99
0.48	18.44	0.99	33.89	1.48	43.06	1.98	47.61	2.48	49.34	2.98	49.86	3.48	49.98	3.98	49.99
0.49	18.79	1.00	34.13	1.49	43.19	1.99	47.67	2.49	49.36	2.99	49.86	3.49	49.98	3.99	49.99

Table 2 *t* Values Needed for Rejection of the Null Hypothesis

How to use this table:

1. Compute the *t* value test statistic.
2. Compare the obtained *t* value to the critical value listed in this table. Be sure you have calculated the number of degrees of freedom correctly and you have selected an appropriate level of significance.
3. If the obtained value is greater than the critical or tabled value, the null hypothesis (that the means are equal) is not the most attractive explanation for any observed differences.
4. If the obtained value is less than the critical or table value, the null hypothesis is the most attractive explanation for any observed differences.

<i>One-Tailed Test</i>				<i>Two-Tailed Test</i>			
<i>df</i>	0.10	0.05	0.01	<i>df</i>	0.10	0.05	0.01
1	3.078	6.314	31.821	1	6.314	12.706	63.657
2	1.886	2.92	6.965	2	2.92	4.303	9.925
3	1.638	2.353	4.541	3	2.353	3.182	5.841
4	1.533	2.132	3.747	4	2.132	2.776	4.604
5	1.476	2.015	3.365	5	2.015	2.571	4.032
6	1.44	1.943	3.143	6	1.943	2.447	3.708
7	1.415	1.895	2.998	7	1.895	2.365	3.5
8	1.397	1.86	2.897	8	1.86	2.306	3.356
9	1.383	1.833	2.822	9	1.833	2.262	3.25
10	1.372	1.813	2.764	10	1.813	2.228	3.17
11	1.364	1.796	2.718	11	1.796	2.201	3.106
12	1.356	1.783	2.681	12	1.783	2.179	3.055
13	1.35	1.771	2.651	13	1.771	2.161	3.013
14	1.345	1.762	2.625	14	1.762	2.145	2.977
15	1.341	1.753	2.603	15	1.753	2.132	2.947
16	1.337	1.746	2.584	16	1.746	2.12	2.921
17	1.334	1.74	2.567	17	1.74	2.11	2.898
18	1.331	1.734	2.553	18	1.734	2.101	2.879
19	1.328	1.729	2.54	19	1.729	2.093	2.861
20	1.326	1.725	2.528	20	1.725	2.086	2.846
21	1.323	1.721	2.518	21	1.721	2.08	2.832
22	1.321	1.717	2.509	22	1.717	2.074	2.819
23	1.32	1.714	2.5	23	1.714	2.069	2.808
24	1.318	1.711	2.492	24	1.711	2.064	2.797
25	1.317	1.708	2.485	25	1.708	2.06	2.788
26	1.315	1.706	2.479	26	1.706	2.056	2.779
27	1.314	1.704	2.473	27	1.704	2.052	2.771

Table 2 (Continued)

<i>One-Tailed Test</i>				<i>Two-Tailed Test</i>			
<i>df</i>	<i>0.10</i>	<i>0.05</i>	<i>0.01</i>	<i>df</i>	<i>0.10</i>	<i>0.05</i>	<i>0.01</i>
28	1.313	1.701	2.467	28	1.701	2.049	2.764
29	1.312	1.699	2.462	29	1.699	2.045	2.757
30	1.311	1.698	2.458	30	1.698	2.043	2.75
35	1.306	1.69	2.438	35	1.69	2.03	2.724
40	1.303	1.684	2.424	40	1.684	2.021	2.705
45	1.301	1.68	2.412	45	1.68	2.014	2.69
50	1.299	1.676	2.404	50	1.676	2.009	2.678
55	1.297	1.673	2.396	55	1.673	2.004	2.668
60	1.296	1.671	2.39	60	1.671	2.001	2.661
65	1.295	1.669	2.385	65	1.669	1.997	2.654
70	1.294	1.667	2.381	70	1.667	1.995	2.648
75	1.293	1.666	2.377	75	1.666	1.992	2.643
80	1.292	1.664	2.374	80	1.664	1.99	2.639
85	1.292	1.663	2.371	85	1.663	1.989	2.635
90	1.291	1.662	2.369	90	1.662	1.987	2.632
95	1.291	1.661	2.366	95	1.661	1.986	2.629
100	1.29	1.66	2.364	100	1.66	1.984	2.626
Infinity	1.282	1.645	2.327	Infinity	1.645	1.96	2.576

Table 3 Critical Values for Analysis of Variance or *F* Test

How to use this table:

1. Compute the *F* value.
2. Determine the number of degrees of freedom for the numerator ($k - 1$) and the number of degrees of freedom for the denominator ($n - k$).
3. Locate the critical value by reading across to locate the degrees of freedom in the numerator and down to locate the degrees of freedom in the denominator. The critical value is at the intersection of this column and row.
4. If the obtained value is greater than the critical or tabled value, the null hypothesis (that the means are equal to one another) is not the most attractive explanation for any observed differences.
5. If the obtained value is less than the critical or tabled value, the null hypothesis is the most attractive explanation for any observed differences.

<i>df for the Denominator</i>	<i>Type I Error Rate</i>	<i>df for the Numerator</i>					
		1	2	3	4	5	6
1	.01	4052.00	4999.00	5403.00	5625.00	5764.00	5859.00
	.05	162.00	200.00	216.00	225.00	230.00	234.00
	.10	39.90	49.50	53.60	55.80	57.20	58.20
2	.01	98.50	99.00	99.17	99.25	99.30	99.33
	.05	18.51	19.00	19.17	19.25	19.30	19.33
	.10	8.53	9.00	9.16	9.24	9.29	9.33
3	.01	34.12	30.82	29.46	28.71	28.24	27.91
	.05	10.13	9.55	9.28	9.12	9.01	8.94
	.10	5.54	5.46	5.39	5.34	5.31	5.28
4	.01	21.20	18.00	16.70	15.98	15.52	15.21
	.05	7.71	6.95	6.59	6.39	6.26	6.16
	.10	.55	4.33	4.19	4.11	4.05	4.01
5	.01	16.26	13.27	12.06	11.39	10.97	10.67
	.05	6.61	5.79	5.41	5.19	5.05	4.95
	.10	4.06	3.78	3.62	3.52	3.45	3.41
6	.01	13.75	10.93	9.78	9.15	8.75	8.47
	.05	5.99	5.14	4.76	4.53	4.39	4.28
	.10	3.78	3.46	3.29	3.18	3.11	3.06
7	.01	12.25	9.55	8.45	7.85	7.46	7.19
	.05	5.59	4.74	4.35	4.12	3.97	3.87
	.10	3.59	3.26	3.08	2.96	2.88	2.83
8	.01	11.26	8.65	7.59	7.01	6.63	6.37
	.05	5.32	4.46	4.07	3.84	3.69	3.58
	.10	3.46	3.11	2.92	2.81	2.73	2.67

Table 3 (Continued)

<i>df for the Denominator</i>	<i>Type I Error Rate</i>	<i>df for the Numerator</i>					
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
9	.01	10.56	8.02	6.99	6.42	6.06	5.80
	.05	5.12	4.26	3.86	3.63	3.48	3.37
	.10	3.36	3.01	2.81	2.69	2.61	2.55
10	.01	10.05	7.56	6.55	6.00	5.64	5.39
	.05	4.97	4.10	3.71	3.48	3.33	3.22
	.10	3.29	2.93	2.73	2.61	2.52	2.46
11	.01	9.65	7.21	6.22	5.67	5.32	5.07
	.05	4.85	3.98	3.59	3.36	3.20	3.10
	.10	3.23	2.86	2.66	2.54	2.45	2.39
12	.01	9.33	6.93	5.95	5.41	5.07	4.82
	.05	4.75	3.89	3.49	3.26	3.11	3.00
	.10	3.18	2.81	2.61	2.48	2.40	2.33
13	.01	9.07	6.70	5.74	5.21	4.86	4.62
	.05	4.67	3.81	3.41	3.18	3.03	2.92
	.10	3.14	2.76	2.56	2.43	2.35	2.28
14	.01	8.86	6.52	5.56	5.04	4.70	4.46
	.05	4.60	3.74	3.34	3.11	2.96	2.85
	.10	3.10	2.73	2.52	2.40	2.31	2.24
15	.01	8.68	6.36	5.42	4.89	4.56	4.32
	.05	4.54	3.68	3.29	3.06	2.90	2.79
	.10	3.07	2.70	2.49	2.36	2.27	2.21
16	.01	8.53	6.23	5.29	4.77	4.44	4.20
	.05	4.49	3.63	3.24	3.01	2.85	2.74
	.10	3.05	2.67	2.46	2.33	2.24	2.18
17	.01	8.40	6.11	5.19	4.67	4.34	4.10
	.05	4.45	3.59	3.20	2.97	2.81	2.70
	.10	3.03	2.65	2.44	2.31	2.22	2.15
18	.01	8.29	6.01	5.09	4.58	4.25	4.02
	.05	4.41	3.56	3.16	2.93	2.77	2.66
	.10	3.01	2.62	2.42	2.29	2.20	2.13
19	.01	8.19	5.93	5.01	4.50	4.17	3.94
	.05	4.38	3.52	3.13	2.90	2.74	2.63
	.10	2.99	2.61	2.40	2.27	2.18	2.11
20	.01	8.10	5.85	4.94	4.43	4.10	3.87
	.05	4.35	3.49	3.10	2.87	2.71	2.60
	.10	2.98	2.59	2.38	2.25	2.16	2.09

(Continued)

Table 3 (Continued)

<i>df for the Denominator</i>	<i>Type I Error Rate</i>	<i>df for the Numerator</i>					
		1	2	3	4	5	6
21	.01	8.02	5.78	4.88	4.37	4.04	3.81
	.05	4.33	3.47	3.07	2.84	2.69	2.57
	.10	2.96	2.58	2.37	2.23	2.14	2.08
22	.01	7.95	5.72	4.82	4.31	3.99	3.76
	.05	4.30	3.44	3.05	2.82	2.66	2.55
	.10	2.95	2.56	2.35	2.22	2.13	2.06
23	.01	7.88	5.66	4.77	4.26	3.94	3.71
	.05	4.28	3.42	3.03	2.80	2.64	2.53
	.10	2.94	2.55	2.34	2.21	2.12	2.05
24	.01	7.82	5.61	4.72	4.22	3.90	3.67
	.05	4.26	3.40	3.01	2.78	2.62	2.51
	.10	2.93	2.54	2.33	2.20	2.10	2.04
25	.01	7.77	5.57	4.68	4.18	3.86	3.63
	.05	4.24	3.39	2.99	2.76	2.60	2.49
	.10	2.92	2.53	2.32	2.19	2.09	2.03
26	.01	7.72	5.53	4.64	4.14	3.82	3.59
	.05	4.23	3.37	2.98	2.74	2.59	2.48
	.10	2.91	2.52	2.31	2.18	2.08	2.01
27	.01	7.68	5.49	4.60	4.11	3.79	3.56
	.05	4.21	3.36	2.96	2.73	2.57	2.46
	.10	2.90	2.51	2.30	2.17	2.07	2.01
28	.01	7.64	5.45	4.57	4.08	3.75	3.53
	.05	4.20	3.34	2.95	2.72	2.56	2.45
	.10	2.89	2.50	2.29	2.16	2.07	2.00
29	.01	7.60	5.42	4.54	4.05	3.73	3.50
	.05	4.18	3.33	2.94	2.70	2.55	2.43
	.10	2.89	2.50	2.28	2.15	2.06	1.99
30	.01	7.56	5.39	4.51	4.02	3.70	3.47
	.05	4.17	3.32	2.92	2.69	2.53	2.42
	.10	2.88	2.49	2.28	2.14	2.05	1.98
35	.01	7.42	5.27	4.40	3.91	3.59	3.37
	.05	4.12	3.27	2.88	2.64	2.49	2.37
	.10	2.86	2.46	2.25	2.14	2.02	1.95
40	.01	7.32	5.18	4.31	3.91	3.51	3.29
	.05	4.09	3.23	2.84	2.64	2.45	2.34
	.10	2.84	2.44	2.23	2.11	2.00	1.93
45	.01	7.23	5.11	4.25	3.83	3.46	3.23
	.05	4.06	3.21	2.81	2.61	2.42	2.31
	.10	2.82	2.43	2.21	2.09	1.98	1.91

Table 3 (Continued)

<i>df for the Denominator</i>	<i>Type I Error Rate</i>	<i>df for the Numerator</i>					
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
50	.01	7.17	5.06	4.20	3.77	3.41	3.19
	.05	4.04	3.18	2.79	2.58	2.40	2.29
	.10	2.81	2.41	2.20	2.08	1.97	1.90
55	.01	7.12	5.01	4.16	3.72	3.37	3.15
	.05	4.02	3.17	2.77	2.56	2.38	2.27
	.10	2.80	2.40	2.19	2.06	1.96	1.89
60	.01	7.08	4.98	4.13	3.68	3.34	3.12
	.05	4.00	3.15	2.76	2.54	2.37	2.26
	.10	2.79	2.39	2.18	2.05	1.95	1.88
65	.01	7.04	4.95	4.10	3.65	3.31	3.09
	.05	3.99	3.14	2.75	2.53	2.36	2.24
	.10	2.79	2.39	2.17	2.04	1.94	1.87
70	.01	7.01	4.92	4.08	3.62	3.29	3.07
	.05	3.98	3.13	2.74	2.51	2.35	2.23
	.10	2.78	2.38	2.16	2.03	1.93	1.86
75	.01	6.99	4.90	4.06	3.60	3.27	3.05
	.05	3.97	3.12	2.73	2.50	2.34	2.22
	.10	2.77	2.38	2.16	2.03	1.93	1.86
80	.01	3.96	4.88	4.04	3.56	3.26	3.04
	.05	6.96	3.11	2.72	2.49	2.33	2.22
	.10	2.77	2.37	2.15	2.02	1.92	1.85
85	.01	6.94	4.86	4.02	3.55	3.24	3.02
	.05	3.95	3.10	2.71	2.48	2.32	2.21
	.10	2.77	2.37	2.15	2.01	1.92	1.85
90	.01	6.93	4.85	4.02	3.54	3.23	3.01
	.05	3.95	3.10	2.71	2.47	2.32	2.20
	.10	2.76	2.36	2.15	2.01	1.91	1.84
95	.01	6.91	4.84	4.00	3.52	3.22	3.00
	.05	3.94	3.09	2.70	2.47	2.31	2.20
	.10	2.76	2.36	2.14	2.01	1.91	1.84
100	.01	6.90	4.82	3.98	3.51	3.21	2.99
	.05	3.94	3.09	2.70	2.46	2.31	2.19
	.10	2.76	2.36	2.14	2.00	1.91	1.83
Infinity	.01	6.64	4.61	3.78	3.32	3.02	2.80
	.05	3.84	3.00	2.61	2.37	2.22	2.10
	.10	2.71	2.30	2.08	1.95	1.85	1.78

Table 4 Values of the Correlation Coefficient Needed for Rejection of the Null Hypothesis

How to use this table:

1. Compute the value of the correlation coefficient.
2. Compare the value of the correlation coefficient with the critical value listed in this table.
3. If the obtained value is greater than the critical or tabled value, the null hypothesis (that the correlation coefficient is equal to 0) is not the most attractive explanation for any observed differences.
4. If the obtained value is less than the critical or tabled value, the null hypothesis is the most attractive explanation for any observed differences.

<i>One-Tailed Test</i>			<i>Two-Tailed Test</i>		
<i>df</i>	.05	.01	<i>df</i>	.05	.01
1	.9877	.9995	1	.9969	.9999
2	.9000	.9800	2	.9500	.9900
3	.8054	.9343	3	.8783	.9587
4	.7293	.8822	4	.8114	.9172
5	.6694	.832	5	.7545	.8745
6	.6215	.7887	6	.7067	.8343
7	.5822	.7498	7	.6664	.7977
8	.5494	.7155	8	.6319	.7646
9	.5214	.6851	9	.6021	.7348
10	.4973	.6581	10	.5760	.7079
11	.4762	.6339	11	.5529	.6835
12	.4575	.6120	12	.5324	.6614
13	.4409	.5923	13	.5139	.6411
14	.4259	.5742	14	.4973	.6226
15	.4120	.5577	15	.4821	.6055
16	.4000	.5425	16	.4683	.5897
17	.3887	.5285	17	.4555	.5751
18	.3783	.5155	18	.4438	.5614
19	.3687	.5034	19	.4329	.5487
20	.3598	.4921	20	.4227	.5368
25	.3233	.4451	25	.3809	.4869
30	.2960	.4093	30	.3494	.4487
35	.2746	.3810	35	.3246	.4182
40	.2573	.3578	40	.3044	.3932
45	.2428	.3384	45	.2875	.3721
50	.2306	.3218	50	.2732	.3541
60	.2108	.2948	60	.2500	.3248
70	.1954	.2737	70	.2319	.3017
80	.1829	.2565	80	.2172	.2830
90	.1726	.2422	90	.2050	.2673
100	.1638	.2301	100	.1946	.2540

Table 5 Critical Values for the Chi-Square Test

How to use this table:

1. Compute the χ^2 value.
2. Determine the number of degrees of freedom for the rows ($R - 1$) and the number of degrees of freedom for the columns ($C - 1$). If it's a one-dimensional table, then you have only columns.
3. Locate the critical value by locating the degrees of freedom in the titled (df) column, and under the appropriate column for level of significance, read across.
4. If the obtained value is greater than the critical or tabled value, the null hypothesis (that the frequencies are equal to one another) is not the most attractive explanation for any observed differences.
5. If the obtained value is less than the critical or tabled value, the null hypothesis is the most attractive explanation for any observed differences.

<i>df</i>	<i>Level of Significance</i>		
	<i>.10</i>	<i>.05</i>	<i>.01</i>
1	2.71	3.84	6.64
2	4.00	5.99	9.21
3	6.25	7.82	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	16.99	18.31	23.21
11	17.28	19.68	24.72
12	18.65	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.60	33.41
18	25.99	28.87	34.80
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.81
26	35.56	38.88	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

Appendix B

Internet Sites About Statistics

What follows is a listing of Internet sites and a brief description of each that focus on the general areas of statistics and measurement. Also included are sites where data (on many different topics) have been collected and can be accessed.

As you use these, keep in mind the following:

- Internet addresses (known as URLs) often change, as does the content. If one of these Internet addresses does not work, search for the name of the site using any search engine.
- Any Internet site is only as good as its content. For example, N or $N - 1$ might be given as the correct denominator for a formula, and although that might be true, you should double check any information with another Internet resource or a book on the subject.
- If you find something that is inaccurate on a site, contact the Webmaster or the author of the site and let him or her know that a correction needs to be made.

Name: statistics.com

Where to find it: <http://www.statistics.com/>

If there is a queen of statistics sites, then statistics.com is it. It offers not only links to hundreds of other sites and an online introductory statistics course, but also online professional development courses. You can try statistics software, look at the free stuff available on the Web, get help if you're a teacher with quizzes and other teaching materials, and

even participate in online discussions. This is *the* place to start your travels.

Name: U.S. Department of Labor, Bureau of Labor Statistics

Where to find it: <http://www.bls.gov/>

Local, state, and federal government agencies are data warehouses, full of information about everything from employment to demographics to consumer spending. This particular site (which is relatively old at 10 years on the Web) is for the Bureau of Labor Statistics, the principal fact-finding agency for the federal government in the areas of labor economics and statistics. It is full of numbers and ideas. Some of the data can be downloaded as HTML or Excel files, and you can also get historical data going back 10 years in some instances.

Name: Probability and Quintile Applets

Where to find it: <http://www.stat.stanford.edu/~naras/jsm/FindProbability.html>

Applets are small programs that can visually represent an idea or a process very effectively. These two, by Balasubramanian Narasimhan from Stanford University, do such things as compute the probability of a score under the normal curve (see Figure 1 on the following page) and calculate the quintiles (fifths) of a distribution. They are easy to use, fun to play with, and very instructional. You can find another similar applet by Gary McClelland at <http://psych.colorado.edu/~mcclella/java/normal/handleNormal.html>

Probability Applet

The first applet below calculates areas under the standard normal density curves. You are presented with two fields and three buttons on top of a normal density curve. The buttons allow you to calculate areas to the left or to the right or between.

- For areas to the left, use the left data entry field.
- For areas to the right, use the right data entry field.
- For areas between two points, use both fields.

The interface is quite forgiving. If you mess up, it takes you right to the field where you have to type.

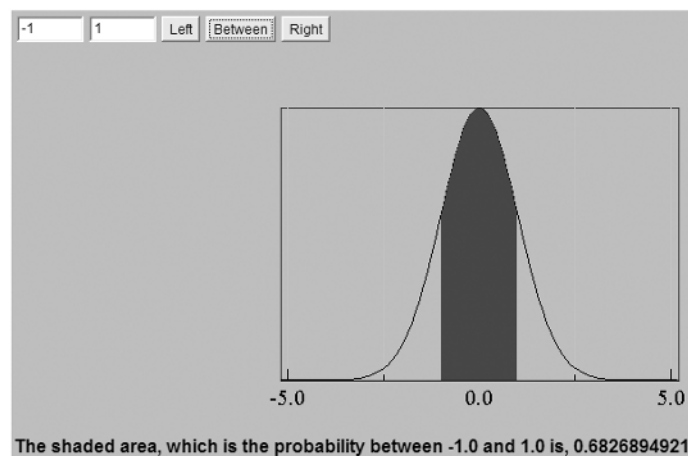


Figure 1 Probability Applet

Name: FedStats

Where to find it: <http://www.fedstats.gov/>

Here's another huge storehouse of data that is the entry point for many different federal agencies. You can easily access data from individual states or from agencies by subjects (such as health), access published collections of statistics, and even get the kids involved in child-oriented agency Web sites both entertaining and educational.

Name: Random Birthday Applet

Where to find it: <http://www-stat.stanford.edu/~susan/surprise/Birthday.html>

This is an incredible illustration of how probability works. You enter the number of birthdays you want generated at random, and the laws of probability should operate so that in a group of 30 such random selections, the odds are very high that there will be at least two matches for the same birthday. Try it—it works.

Name: The Statistics Homepage

Where to find it: <http://www.statsoftinc.com/textbook/stathome.html>

Here you'll find a self-contained course in basic statistics, brought to you by the people who developed and sell StatSoft, one of many statistical programs. On this site, you will find tutorials that take you from the elementary concepts of statistics through the more advanced topics, such as factor and discriminant analysis.

Name: National Center for Health Statistics

Where to find it: <http://www.cdc.gov/nchs/>

The National Center for Health Statistics compiles information that helps guide actions and policies to improve health in the United States. Among other things, these data are used to help identify health problems, evaluate the effectiveness of programs, and provide data for policymakers.

Name: The World Wide Web Virtual Library: Statistics

Where to find it: <http://www.stat.ufl.edu/vlib/statistics.html>

The good people at the University of Florida's Department of Statistics bring you this page, which contains links to statistics departments all over the world. It provides a great deal of information about graduate study in these areas as well as other resources.

Name: Social Statistics Briefing Room

Where to find it: <http://www.whitehouse.gov/fsbr/ssbr.html>

This service, which calls the White House home, provides access to current federal social statistics and links to information from a wide range of federal agencies. This is a very good, and broad, starting point to access data made available through different agencies.

Name: Statistics on the Web

Where to find it: <http://my.execpc.com/~helberg/statistics.html>

More groupings of URLs and Internet addresses from Clay Helberg. A bit like statistics.com, but full of listings of professional organizations, publications, and software packages (many of which you can download for a trial).

Name: Food and Agriculture Organization for the United Nations

Where to find it: <http://faostat.fao.org/>

If you want to go international, this is a site containing online information (in multilingual formats) and databases for more than 3 million time series records covering international statistics in areas such as production, population, and exports.

Name: Web Pages That Perform Statistical Calculations!

Where to find it: <http://members.aol.com/johnp71/javastat.html>

At the time of this writing, this site contains more than 600 links to books, tutorials, free software, and interactive tools, such as a guide to what statistical test to use to answer what questions, all assembled by John Pezzullo.

Name: Free Statistical Software

Where to find it: <http://freestatistics.altervista.org/stat.php>

An extensive collection of statistical analysis software packages that range from simple programs for students to advanced programs that do everything from statistical visualization to time series analysis. Many of these programs are freeware, and many are open source, available to be modified by users.

Name: Java Applets

Where to find it: <http://www.stat.duke.edu/sites/java.html>

The Institute of Statistics and Decision Sciences at Duke University and NWP Associates put together a collection of Java applets (Java is the language in which these small programs are written, and applets

are small applications) that allows the user to demonstrate interactively various statistical techniques and tools, such as constructing histograms and illustrating how the central limit theorem works.

Name: HyperStat Online Textbook

Where to find it: <http://davidmlane.com/hyperstat/>

This site contains an entire online course in basic statistics from David Lane that covers every topic from simple descriptive statistics to effect size. The “Hyper” nature of the site allows the user to easily move from one topic to another through the extensive use of live links. And, as a bonus, each new screen has additional links to sites that focus on learning statistics.

Name: Rice Virtual Lab in Statistics

Where to find it: <http://www.ruf.rice.edu/~lane/rvls.html>

This is where the HyperStat Online Textbook has its home and is the main page (also done by David Lane) of Rice University’s statistics program. In addition to the HyperStat link, it has links to simulations, case studies, and a terrific set of applets that are very useful for teaching and demonstration purposes.

Name: Reliability, Validity, and Fairness of Classroom Assessments

Where to find it: <http://www.ncrel.org/sdrs/areas/issues/methods/assment/as5relia.htm>

A discussion of the reliability, validity, and fairness of classroom testing from the North Central Educational Laboratory.

Name: The Multitrait Multimethod Matrix

Where to find it: <http://www.socialresearchmethods.net/kb/mtmmmat.htm>

A very good site for a discussion of validity issues in measurement in general and specific discussion about the multitrait multimethod brought to you by William M. K. Trochim.

Name: Content Validity, Face Validity, and Quantitative Face Validity

Where to find it: <http://www.burns.com/wcbcontval.htm>

Although a bit dated (around 1996), this Web site offers a detailed discussion by William C. Burns on content, face, quantitative, and other types of validity.

Name: The National Education Association

Where to find it: <http://www.nea.org/parents/testing/guide.html>.also

This national organization of teaching professionals provides assistance to parents, teachers, and others in understanding test scores.

Name: The Learning Center

Where to find it: <http://webster.commnet.edu/faculty/~simonds/tests.htm>

It's a reality that other than through studying, test scores can be improved if test takers understand the different demands of different types of tests. This item contains information on using different strategies to increase test scores.

Name: The Advanced Placement

Where to find it: <http://apbio.biosci.uga.edu/exam/Essays/>

This is an old site, but people at the University of Georgia have posted items from a variety of different topic areas covered in the Advanced Placement (AP) exams that high school students can take in a step to qualify for college credit.

Name: Essay Question

Where to find it: http://www.salon.com/tech/feature/1999/05/25/computer_grading/

Salon.com offers a discussion of automated grading in general and specially, as well as essay question grading using computers.

Name: Matching Questions on Minerals and Rocks

Where to find it: <http://www.usd.edu/esci/exams/matching.html>

A good example of how easy it is to adapt matching questions to an interactive electronic format.

An increasingly large part of doing research, as well as other intensive, more qualitative projects, involves specially designed software. At <http://www.scolari.com/>, you can find a listing of several different types and explore which might be right for you if you intend to pursue this method (interviewing) and this methodology (qualitative).

FairTest—The National Center for Fair and Open Testing at <http://www.fairtest.org/index.htm> has as its mission to “end the misuses and flaws of standardized testing and to ensure that evaluation of students, teachers and schools is fair, open, valid and educationally beneficial.” A really interesting site to visit.

Preparing Students to Take Standardized Achievement Tests (at <http://pareonline.net/getvn.asp?v=1&n=11>) was written by William A. Mehrens (and first appeared in *Practical Assessment, Research & Evaluation*) for school administrators and teachers and discusses what test scores mean and how they can be most useful in understanding children's performance.

The Clifton StrengthsFinder™ at <http://education.gallup.com/content/default.asp?ci=886> is a Web-based assessment tool published by the Gallup Organization (yep, the poll people) to help people better understand their talents and strengths by measuring the presence of 34 themes of talent. You might want to take it and explore these themes.

Find out just about everything you ever wanted to know (and more) about human intelligence at Human Intelligence: Historical Influences, Current Controversies and Teaching Resources at <http://www.indiana.edu/~intell/>

The following text is taken from Neil J. Salkind's best-selling introduction to statistics text, *Statistics for People Who (Think They) Hate Statistics*, 2nd edition (2004).

Pages and pages of every type of statistical resource you can want has been creatively assembled by Professor David W. Stockburger at <http://www.psychstat.smsu.edu/scripts/dws148f/statisticsresourcesmain.asp>. This site receives the gold medal of statistics sites. Don't miss it.

For example, take a look at Berrie's page (at <http://www.huizen.dds.n~berrie/>) and see some QuickTime (short movies) of the effects of changing certain data points on the value of the mean and standard deviation. Or, look at the different home pages that have been created by instructors for courses offered around the country. Or, look at all of the different software packages that can do statistical analysis.

Want to draw a histogram? How about a table of random numbers? A sample-size calculator? The Statistical Calculators page at <http://www.stat.ucla.edu/calculators/> has just about every type (more than 15) of calculator and table you could need. Enough to carry you through any statistics course that you might take and even more.

For example, you can click on the Random Permutations link and complete the two boxes (as you see in Figure 2 for 2 random permutations of 100 integers), and you get the number of permutations you want. This is very handy when you need a table of

random numbers for a specific number of participants so you can assign them to groups.

The History of Statistics page located at <http://www.Anselm.edu/homepage/jpitocch/biostatshist.html> contains portraits and bibliographies of famous statisticians and a time line of important contributions to the field of statistics. So, do names like Bernoulli, Galton, Fisher, and Spearman pique your curiosity? How about the development of the first test between two averages during the early 20th century? It might seem a bit boring until you have a chance to read about the people who make up this field and their ideas—in sum, pretty cool ideas and pretty cool people.

SurfStat Australia (at <http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html>) is the online component of a basic stat course taught at the University of Newcastle, Australia, but has grown far beyond just the notes originally written by Annette Dobson in 1987, and updated over several years' use by Anne Young, Bob Gibberd, and others. Among other things, SurfStat contains a complete interactive statistics text. Besides the text, there are exercises, a list of other statistics sites on the Internet, and a collection of Java applets (cool little programs you can use to work with different statistical procedures).

This online tutorial with 18 lessons, at <http://www.davidmlane.com/hyperstat/index.html>, offers nicely designed and user-friendly coverage of the important basic topics. What we really liked about the site was the glossary, which uses hypertext to connect different concepts to one another. For example, in Figure 3, you can see the definition of descriptive statistics also linked to

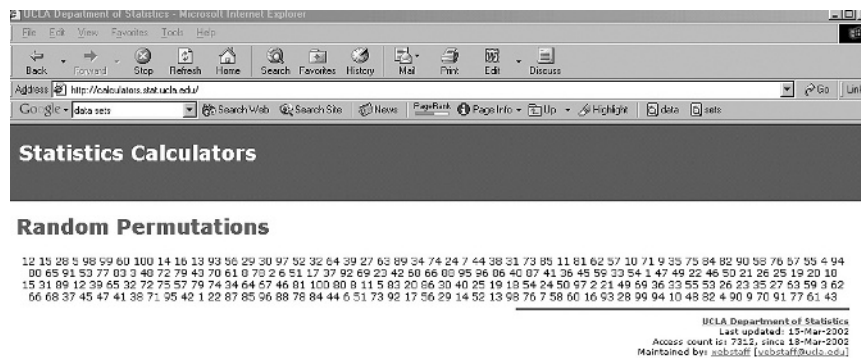


Figure 2 Generating a Set of Random Numbers

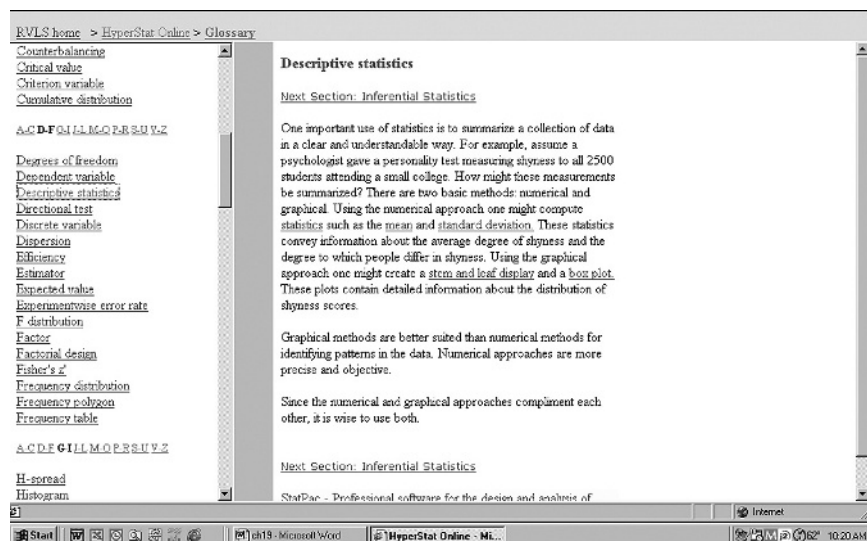


Figure 3 Sample HyperStat Screen

other glossary terms, such as mean, standard deviation, and box plot. Click on any of those and zap! you're there.

There are data all over the place, ripe for the picking. Here are just a few. What to do with these? Download them to be used as examples in your work or as examples of analysis that you might want to do, and you can use these as a model.

- Statistical Reference Datasets at <http://www.itl.nist.gov/div898/strd/>
- United States Census Bureau (a huge collection and a gold mine of data) at http://factfinder.census.gov/servlet/DatasetMainPage-Servlet?_lang=en
- The Data and Story Library (<http://lib.stat.cmu.edu/DASL/>) with great annotations about the data (look for the stories link)
- Tons of economic data sets at Growth Data Sets (at <http://www.bris.ac.uk/Depts/Economics/Growth/datasets.htm>)

Then there are all the data sets that are available through the federal government (besides the census). Your tax money supports it, so why not use it? For example, there's FEDSTATS (at <http://www.fedstats.gov/>), where more than 70 agencies in the U.S. federal government produce statistics of interest to the public. The Federal Interagency Council on Statistical Policy

maintains this site to provide easy access to the full range of statistics and information produced by these agencies for public use. Here you can find country profiles contributed by the (boo!) CIA; public school student, staff, and faculty data (from the National Center for Education Statistics); and the Atlas of the United States Mortality (from the National Center for Health Statistics). What a ton of data!

The University of Michigan's Statistical Resources on the Web (at <http://www.lib.umich.edu/govdocs/stats.html>) has hundreds and hundreds of resource links,

including those to banking, book publishing, the elderly, and, for those of you with allergies, pollen count. Browse, search for what exactly it is that you need—no matter, you are guaranteed to find something interesting.

At <http://mathforum.org/workshops/sum96/data.collections/datalibrary/data.set6.html>, you can find a data set including the 1994 National League Baseball Salaries or the data on TV, Physicians, and Life Expectancy. Nothing earth-shaking, just fun to download and play with.

The World Wide Web Virtual Library: Statistics is the name of the page, but the one-word title is misleading because the site (from the good people at the University of Florida at <http://www.stat.ufl.edu/vlib/statistics.html>) includes information on just about every facet of the topic, including data sources, job announcements, departments, divisions and schools of statistics (a huge description of programs all over the world), statistical research groups, institutes and associations, statistical services, statistical archives and resources, statistical software vendors and software, statistical journals, mailing list archives, and related fields. Tons of great information is available here. Make it a stop along the way.

Statistics on the Web at <http://www.maths.uq.edu.au/~gks/webguide/datasets.html> is another location that's

just full of information and references that you can easily access. Here, you'll find information on professional organizations, institutes and consulting groups, educational resources, Web courses, online textbooks, publications and publishers, statistics book lists, software-oriented pages, mailing lists and discussion groups, and even information on statisticians and other statistical people.

If you do ever have to teach statistics, or even tutor fellow students, this is one place you'll want to visit: <http://noppa5.pc.helsinki.fi/links.html>. It contains hundreds of resources on every topic that was covered in *Statistics for People Who (Think They) Hate Statistics* and more. You name it and it's here: regression, demos, history, Sila (a demonstration of inference), an interactive online tutorial, statistical graphics, handouts to courses, teaching materials, journal articles, and even quizzes! Whew, what a deal. There tends to be a lot of material that may not be suited to what you are doing in this class, but this wide net has certainly captured some goodies.



Figure 4 Selecting the Correct Stat Technique to Use—Just a Few Clicks Away

Statistics.com (www.statistics.com) has it all—a wealth of information on courses, software, statistical methods, jobs, books, and even a homework helper. For example, if you want to know about free Web-based stat packages, click on that link on the left-hand side of the page. Here's one (see Figure 4) from Dr. Bill Trochim. . . . You just click your way through answering questions to get the answer to what type of analysis should be used.

Appendix C

Glossary

The following text is taken from Neil J. Salkind's best-selling introduction to statistics text, *Statistics for People Who (Think They) Hate Statistics*, 2nd edition (2004).

Analysis of variance

A test for the difference between two or more means. A simple analysis of variance (or ANOVA) has only one independent variable, whereas a factorial analysis of variance tests the means of more than one independent variable. One-way analysis of variance looks for differences between the means of more than two groups.

Arithmetic mean

A measure of central tendency that sums all the scores in the data sets and divides by the number of scores.

Asymptotic

The quality of the normal curve such that the tails never touch.

Average

The most representative score in a set of scores.

Bell-shaped curve

A distribution of scores that is symmetrical about the mean, median, and mode and has asymptotic tails.

Class interval

The upper and lower boundaries of a set of scores used in the creation of a frequency distribution.

Coefficient of alienation

The amount of variance unaccounted for in the relationship between two variables.

Coefficient of determination

The amount of variance accounted for in the relationship between two variables.

Coefficient of nondetermination

See coefficient of alienation

Concurrent validity

A type of validity that examines how well a test outcome is consistent with a criterion that occurs in the present.

Construct validity

A type of validity that examines how well a test reflects an underlying construct.

Content validity

A type of validity that examines how well a test samples a universe of items.

Correlation coefficient

A numerical index that reflects the relationship between two variables.

Correlation matrix

A set of correlation coefficients.

Criterion

Another term for the outcome variable.

Criterion validity

A type of validity that examines how well a test reflects some criterion that occurs either in the present (concurrent) or in the future (predictive).

Critical value

The value necessary for rejection (or nonacceptance) of the null hypothesis.

Cumulative frequency distribution

A frequency distribution that shows frequencies for class intervals along with the cumulative frequency for each.

Data

A record of an observation or an event such as a test score, a grade in math class, or a response time.

Data point

An observation.

Data set

A set of data points.

Degrees of freedom

A value that is different for different statistical tests and approximates the sample size of number of individual cells in an experimental design.

Dependent variable

The outcome variable or the predicted variable in a regression equation.

Descriptive statistics

Values that describe the characteristics of a sample or population.

Direct correlation

A positive correlation where the values of both variables change in the same direction.

Directional research hypothesis

A research hypothesis that includes a statement of inequality.

Effect size

A measure of the magnitude of a particular outcome.

Error in prediction

The difference between the actual score (Y) and the predicted score ($\bar{Y}$).

Error of estimate

See error in prediction

Error score

The part of a test score that is random and contributes to the unreliability of a test.

Factorial analysis of variance

An analysis of variance with more than one factor or independent variable.

Factorial design

A research design where there is more than one treatment variable.

Frequency distribution

A method for illustrating the distribution of scores within class intervals.

Frequency polygon

A graphical representation of a frequency distribution.

Histogram

A graphical representation of a frequency distribution.

Hypothesis

An if-then statement of conjecture that relates variables to one another.

Independent variable

The treatment variable that is manipulated or the predictor variable in a regression equation.

Indirect correlation

A negative correlation where the values of variables move in opposite directions.

Inferential statistics

Tools that are used to infer the results based on a sample to a population.

Interaction effect

The outcome where the effect of one factor is differentiated across another factor.

Internal consistency reliability

A type of reliability that examines the one-dimensional nature of an assessment tool.

Interrater reliability

A type of reliability that examines the consistency of raters.

Interval level of measurement

A scale of measurement that is characterized by equal distances between points on some underlying continuum.

Kurtosis

The quality of a distribution such that it is flat or peaked.

Leptokurtic

The quality of a normal curve that defines its peakedness.

Line of best fit

The regression line that best fits the actual scores and minimizes the error in prediction.

Linear correlation

A correlation that is best expressed as a straight line.

Main effect

In analysis of variance, when a factor or an independent variable has a significant effect upon the outcome variable.

Mean

A type of average where scores are summed and divided by the number of observations.

Mean deviation

The average deviation for all scores from the mean of a distribution.

Measures of central tendency

The mean, median, and mode.

Median

The point at which 50% of the cases in a distribution fall below and 50% fall above.

Midpoint

The central point in a class interval.

Mode

The most frequently occurring score in a distribution.

Multiple regression

A statistical technique where several variables are used to predict one.

Nominal level of measurement

A scale of measurement that is characterized by categories with no order or difference in magnitude.

Nondirectional research hypothesis

A hypothesis that posits no direction, but a difference.

Nonparametric statistics

Distribution-free statistics.

Normal curve

See bell-shaped curve

Null hypothesis

A statement of equality between a set of variables.

Observed score

The score that is recorded or observed.

Obtained value

The value that results from the application of a statistical test.

Ogive

A visual representation of a cumulative frequency distribution.

One-tailed test

A directional test.

One-way analysis of variance

See analysis of variance

Ordinal level of measurement

A scale of measurement that is characterized by an underlying continuum that is ordered.

Outliers

Those scores in a distribution that are noticeably much more extreme than the majority of scores. Exactly what score is an outlier is usually an arbitrary decision made by the researcher.

Parallel forms reliability

A type of reliability that examines the consistency across different forms of the same test.

Parametric statistics

Statistics used for the inference from a sample to a population.

Pearson product-moment correlation

See correlation coefficient

Percentile point

The point at or below where a score appears.

Platykurtic

The quality of a normal curve that defines its flatness.

Population

All the possible subjects or cases of interest.

Post hoc

After the fact, referring to tests done to determine the true source of a difference between three or more groups.

Predictive validity

A type of validity that examines how well a test outcome is consistent with a criterion that occurs in the future.

Predictor

The variable that predicts an outcome.

Range

The highest minus the lowest score, and a gross measure of variability. *Exclusive* range is the highest score minus the lowest score. *Inclusive* range is the highest score minus the lowest score plus 1.

Ratio level of measurement

A scale of measurement that is characterized by an absolute zero.

Regression equation

The equation that defines the points and the line that are closest to the actual scores.

Regression line

The line drawn based on the values in the regression equation.

Reliability

The quality of a test such that it is consistent.

Research hypothesis

A statement of inequality between two variables.

Sample

A subset of a population.

Sampling error

The difference between sample and population values.

Scales of measurement

Different ways of categorizing measurement outcomes.

Scattergram, or scatterplot

A plot of paired data points.

Significance level

The risk set by the researcher for rejecting a null hypothesis when it is true.

Simple analysis of variance

See analysis of variance

Skew, or skewness

The quality of a distribution that defines the disproportionate frequency of certain scores. A longer right tail than left corresponds to a smaller number of occurrences at the high end of the distribution; this is a *positively* skewed distribution. A shorter right tail than left corresponds to a larger number of occurrences at the high end of the distribution; this is a *negatively* skewed distribution.

Source table

A listing of sources of variance in an analysis of variance summary table.

Standard deviation

The average deviation from the mean.

Standard error of estimate

A measure of accuracy in prediction.

Standard score

See *z* score

Statistical significance

See significance level

Statistics

A set of tools and techniques used to organize and interpret information.

Test-retest reliability

A type of reliability that examines consistency over time.

Test statistic value

See obtained value

True score

The unobservable part of an observed score that reflects the actual ability or behavior.

Two-tailed test

A test of a nondirectional hypothesis where the direction of the difference is of little importance.

Type I error

The probability of rejecting a null hypothesis when it is true.

Type II error

The probability of accepting a null hypothesis when it is false.

Unbiased estimate

A conservative estimate of a population parameter.

Validity

The quality of a test such that it measures what it says it does.

Variability

The amount of spread or dispersion in a set of scores.

Variance

The square of the standard deviation, and another measure of a distribution's spread or dispersion.

 Y' or Y prime

The predicted Y value.

 z score

A raw score that is adjusted for the mean and standard deviation of the distribution from which the raw score comes.

Master Bibliography

- Aaron T. Beck Web page: <http://mail.med.upenn.edu/~abeck/>
- Abdi, H. (1987). *Introduction au traitement statistique des données expérimentales*. Grenoble, France: Presses Universitaires de Grenoble.
- Abdi, H. (1990). Additive-tree representations. *Lecture Notes in Biomathematics*, 84, 43–59.
- Abdi, H. (2004). Least squares. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage. Retrieved April 11, 2006, from <http://www.utdallas.edu/~herve/Abdi-LeastSquares-pretty.pdf>
- Abdi, H. (2004). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H. (2004). PLS-regression. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.
- Abdi, H., & Valentin, D. (2006). *Mathématiques pour les sciences cognitives* [Mathematics for cognitive sciences]. Grenoble, France: Presses Universitaires de Grenoble.
- Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition.
- Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition.
- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
- Abdi, H., Valentin, D., O'Toole, A. J., & Edelman, B. (2005). DISTATIS: The analysis of multiple distance matrices. *Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition*, pp. 42–47.
- Abedi, J., & Bruno, J. E. (1989). Test-retest reliability of computer based MCW-APM test scoring methods. *Journal of Computer-Based Instruction*, 16, 29–35.
- Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the child behavior checklist and revised child behavior profile*. Burlington: University of Vermont, Department of Psychiatry.
- Adams, S. (1950). Does face validity exist? *Educational and Psychological Measurement*, 10, 320–328.
- Adams, S. J. (2001, October). Projecting the next decade in safety management: A Delphi technique study. *American Society of Safety Engineers*, 32, 26–29.
- Adaptive testing Web site including SLAT: <http://psychology.gatech.edu/cml/Adaptive/slat.htm>
- Aday, L. A. (1996). *Designing and conducting health surveys* (2nd ed.). San Francisco: Jossey-Bass.
- ade4 package for R: <http://pbil.univ-lyon1.fr/R/rplus/ade4dsR.html> (enables you to enter data and compute dissimilarity coefficients, diversity coefficients, the Principal Coordinates Analysis and the double Principal Coordinates Analysis)
- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- AIMSweb Response to Intervention in a three-tiered model: http://www.aimsweb.com/products/aimsweb_rti.htm
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémia Kiadó.
- Alan S. Kaufman and Nadeen L. Kaufman (test developers) biographical information: <http://www.mhhe.com/mayfieldpub/psychtesting/profiles/karfmann.htm>
- Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10, 364–376.
- Alexander Luria biographies: <http://www.marxists.org/archive/luria/comments/bio.htm> and http://en.wikipedia.org/wiki/Alexander_Luria

- Alexander, H. W. (1961). *Elements of mathematical statistics*. New York: Wiley.
- Alfred Binet and his contributions: <http://www.indiana.edu/~intell/binet.shtml>
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to test theory*. Monterey, CA: Brooks-Cole.
- Alvey, W., & Jamerson, B. (1997). *Record linkage techniques—1997, Proceedings of an International Workshop and Exposition*. Federal Committee on Statistical Methodology, Office of Management and Budget. Retrieved from http://www.fcsm.gov/working-papers/RLT_1997.html
- American Academy of Pediatrics. (2000). Clinical practice guideline: Diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. *Pediatrics*, 105I, 1158–1170.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: Authors.
- American Educational Research Association. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: Author.
- American Guidance Service: <http://www.agsnet.com>
- American Medical Association. (2004). *Code of medical ethics: Current opinions with annotations, 2004–2005*. Chicago: Author.
- American National Standards Institute (ANSI)/American Society for Quality (ASQ). (2003). *Sampling procedures and tables for inspection by attributes*, ANSI/ASQ Z1.4. Milwaukee, WI: ASQ Quality Press.
- American National Standards Institute (ANSI)/American Society for Quality (ASQ). (2003). *Sampling procedures and tables for inspection by variables for percent nonconforming*, ANSI/ASQ Z1.9. Milwaukee, WI: ASQ Quality Press.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060–1073.
- American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.
- American Statistical Association Web site: www.amstat.org
- Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101 et seq.
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). Retrieved July 18, 2005, from <http://epaa.asu.edu/epaa/v10n18>
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Anastasi, A., & Urbina, S. (1997). Part four: Personality testing. In *Psychological testing* (8th ed., pp. 348–471). Upper Saddle River, NJ: Prentice Hall.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice Hall.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.
- Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Erlbaum.
- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.
- Anderson, T., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20, 46–63.
- Andrew, D. M., Paterson, D. G., & Longstaff, H. P. (1979). *Manual for the Minnesota Clerical Test*. New York: Psychological Corporation.
- Andrew L. Comrey: <http://www.today.ucla.edu/2002/020312comrey.html>
- Andrich, D. A., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17(3), 253–276.
- APA Ethics Code (2002): <http://www.apa.org/ethics>
- APA Ethics Code, including a section on assessment: <http://www.apa.org/ethics/code2002.html>
- APA policy statement: <http://www.apa.org/practice/ebpstatement.pdf>
- APA report: <http://www.apa.org/practice/ebpreport.pdf>
- APA statement on test security in educational settings: <http://www.apa.org/science/securetests.html>
- APA Web site: www.apa.org
- The Apoala project is an offspring of GeoVista and is a dynamic parallel coordinate plot, implemented in TCL, designed to show the relationships between multiple variables in large data sets: <http://www.geovista.psu.edu/products/demos/edsall/Tclets072799/pcpdescription.html>
- Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27, 186–190.
- Area under the curve and receiving operator characteristic curve description: <http://www.anaesthetist.com/mnm/stats/roc/>
- Armed Services Vocational Aptitude Battery: <http://www.asvabprogram.com/>
- Arnold, H. J., & Evans, M. G. (1979). Testing multiplicative models does not require ratio scales. *Organizational Behavior and Human Performance*, 24, 214–224.

- Arnold, R., & Lidz, C. (1995). Informed consent: Clinical aspects of consent in health care. In W. Reich (Ed.), *Encyclopedia of bioethics* (Vol. 3, Rev. ed., pp. 1250–1256). New York: Macmillan.
- Aron, A., Aron, E. N., & Coups, E. (2005). *Statistics for psychology* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Artificial neural networks technology: http://www.dacs.dtic.mil/techs/neural/neural_ToC.html
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6, 128–143.
- Association for Psychological Science Web site: www.psychologicalscience.org
- Averages and deviations: http://www.sciencebyjones.com/average_deviation.htm
- Babbage's machines: <http://www.sciencemuseum.org.uk/on-line/babbage/index.asp>
- Bailey, B. J. R. (1977). Tables of the Bonferroni *t* statistic. *Journal of the American Statistical Association*, 72, 469–478.
- Baines, L. S., Joseph, J. T., & Jindal, R. M. (2004). Prospective randomized study of individual and group psychotherapy versus controls in recipients of renal transplants. *Kidney International*, 65, 1523–1755.
- Bak, J. J., & Siperstein, G. N. (1987). Similarity as a factor effecting change in children's attitudes toward mentally retarded peers. *American Journal of Mental Deficiency*, 91(5), 524–531.
- Baker, E. L., & Linn, R. L. (2002). *Validity issues for accountability systems*. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, 48(12), 1210–1218.
- Ball, W. W. R. (1960). *A short account of the history of mathematics*. New York: Dover.
- Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-event system simulation* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Barchard, K. A., & Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational & Psychological Measurement*, 57, 893–905.
- Barchard, K. A., & Hakstian, A. R. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Research*, 32, 169–191.
- Barlow, D. H., & Hersen, M. (1984). *Single case experimental design: Strategies for studying behavior change* (2nd ed.). Needham Heights, MA: Allyn & Bacon.
- Barnard, J., & Rubin, D. B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively-worded stems. *Educational and Psychological Measurement*, 60(3), 361–370.
- Barnsley, M. G. (1988). *Fractals everywhere*. New York: Academic Press.
- Barton, J., & Collins, A. (Eds.). (1997). *Portfolio assessment: A handbook for educators*. Menlo Park, CA: Addison-Wesley.
- Barzilai, J. (1998). On the decomposition of value functions. *Operations Research Letters*, 22, 159–170.
- Barzilai, J. (2001). Notes on the analytic hierarchy process. *Proceedings of the NSF Design and Manufacturing Research Conference* (pp. 1–6). Tampa, Florida.
- Barzilai, J. (2004). Notes on utility theory. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1000–1005.
- Barzilai, J. (2005). Measurement and preference function modelling. *International Transactions in Operational Research*, 12, 173–183.
- Barzilai, J. (2006). Preference modelling in engineering design. In K. Lewis, W. Chen, & L. Schmidt (Eds.), *Decision making in engineering design*. New York: ASME Press.
- BASC-2 information: <http://www.agsnet.com/Group.asp?nGroupInfoID=a30000>
- Basic Personality Inventory: <http://www.sigmaassessmentsystems.com/assessments/bpi.asp>
- Basic vs. applied research (discussion from the Ethical, Legal, and Social Issues in Science site at the University of California): <http://www.lbl.gov/Education/ELSI/research-main.html>
- Basu, A. P., & Dhar, S. K. (1995). Bivariate geometric distribution. *Journal of Applied Statistical Science*, 2, 33–44.
- Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.
- Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, 62, 254–263.
- Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, 45, 1289–1298.

- Beail, N. (2003). Utility of the Vineland Adaptive Behavior Scales in diagnosis and research with adults who have mental retardation. *Mental Retardation*, 41(4), 286–289.
- Bean, P., Loomis, C., Timmel, P., Hallinan, P., Moore, S., Mammel, J., et al. (2004). Outcome variables for anorexic males and females one year after discharge from residential treatment. *Journal of Addictive Diseases*, 23, 83–94.
- Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis*, 42, 569–593.
- Beauchamp, T., & Childress, J. (1994). *Principles of biomedical ethics* (4th ed.). New York: Oxford University Press.
- Becker, R. A., & Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29, 127–142.
- Becker, R. A., Cleveland, W. S., & Weil, G. (1988). The use of brushing and rotation for data analysis. In W. S. Cleveland & M. E. McGill (Eds.), *Dynamic graphics for statistics* (pp. 247–275). Pacific Grove, CA: Wadsworth.
- Bedeian, A. G., Armenakis, A. A., & Randolph, W. A. (1988). The significance of congruence coefficients: A comment and statistical test. *Journal of Management*, 14, 559–566.
- Bedrick, E. J. (1992). A comparison of modified and generalized sample biserial correlation estimators. *Psychometrika*, 57, 183–201.
- Beecher, H. (1966). Special article: Ethics and clinical research. *New England Journal of Medicine*, 274, 1354–1360.
- Belin, T. R., Ishwaran, H., Duan, N., Berry, S., & Kanouse, D. (2004). Identifying likely duplicates by record linkage in a survey of prostitutes. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. New York: Wiley.
- Bell, N. L., & Nagle, R. J. (1999). Interpretive issues with the Roberts Apperception Test for Children: Limitations of the standardization group. *Psychology in the Schools*, 36, 277–283.
- Bell, S. K., & Morgan, S. B. (2000). Children's attitudes and behavioral intentions toward a peer presented as obese: Does a medical explanation for the obesity make a difference? *Journal of Pediatric Psychology*, 25(3), 137–145.
- Bellman, R. E. (1961). *Adaptive control processes*. Princeton, NJ: Princeton University Press.
- The Belmont Report: [http://www.hhs.gov/ohrp/human subjects/guidance/belmont.htm](http://www.hhs.gov/ohrp/human_subjects/guidance/belmont.htm)
- Belmont Report. (1979). *Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education, and Welfare.
- Benjamini, Y. (1988). Opening the box of a boxplot. *American Statistician*, 42, 257–262.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.
- Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (RR-90-7). Princeton, NJ: Educational Testing Service.
- Benson, J., & Wilcox, S. (1981, April). *The effect of positive and negative item phrasing on the measurement of attitudes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles. (ERIC Document Reproduction Service No. ED204404)
- Benzécri, J. P. (1973). *Analyse des données*. Paris: Dunod.
- Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 4, 377–378.
- Berenson, M. L. (1982). Some useful nonparametric tests for ordered alternatives in randomized block experiments. *Communications in Statistics: Theory and Methods*, 11, 1681–1693.
- Berenson, M. L. (1982). A study of several useful nonparametric tests for ordered alternatives in randomized block experiments. *Communications in Statistics: Simulation and Computation*, 11, 563–581.
- Berger, R., & Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11, 283–319.
- Bergus, G. R., & Hamm, R. (1995). Clinical practice: How physicians make medical decisions and why medical decision making can help. *Primary Care*, 22(2), 167–180.
- Berk, R. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.
- Berliner, D. C. (1993, April). Mythology and the American system of education. *Phi Delta Kappan*, pp. 632–640.
- Bermuda Atlantic Time Series Study: <http://coexploration.org/bbsr/classroombats/html/visualization.html> (illustrates use of contour plots in oceanography)
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, UK: Wiley.
- Berndt, E. R. (1991). *The practice of econometrics*. New York: Addison-Wesley.
- Berrin-Wasserman, S., Winnick, W. A., & Borod, J. C. (2003). Effects of stimulus emotionality and sentence generation on memory for words in adults with unilateral brain damage. *Neuropsychology*, 17, 429–438.
- Berry, M., & Linoff, G. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed.). New York: Wiley.

- Bersoff, D. N. (Ed.). (2003). *Ethical conflicts in psychology* (3rd ed.). Washington, DC: American Psychological Association.
- Besag, J. (2001). *Markov Chain Monte Carlo for statistical inference*. Center for Statistics and the Social Sciences Working Paper No. 9. Available at <http://www.csss.washington.edu/Papers/>
- Bettelheim, B. (1967). *The empty fortress*. New York: Free Press.
- Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. Maple Grove, MN: JAM.
- Binet, A., Simon, T., & Terman, L. M. (1980). *The development of intelligence in children* (1916 limited ed.). Nashville, TN: Williams Printing.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.
- Black, M. (1999). *Essentials of Bayley Scales of Infant Development II assessment*. New York: Wiley.
- Blair, R. C., & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistics*, 1, 139–142.
- Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5(4), 309–335.
- Blalock, H. M. (1964). *Causal inference in nonexperimental research*. Chapel Hill: University of North Carolina Press.
- Blank, M., Rose, S. A., & Berlin, L. J. (1978). *The language of learning: The preschool years*. New York: Grune & Stratton.
- BLAST resources: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html>, <http://www.ncbi.nlm.nih.gov/>
- Bloomfield, P. (2000). *Fourier analysis of time series: An introduction*. New York: Wiley.
- Board of Education of the Hendrick Hudson Central School District v. Rowley, 458 U.S. 176 (1982).
- Bockian, N., Meager, S., & Millon, T. (2000). Assessing personality with the Millon Behavioral Health Inventory, the Millon Behavioral Medicine Diagnostic, and the Millon Clinical Multiaxial Inventory. In R. J. Gatchel & J. N. Weisberg (Eds.), *Personality characteristics of patients with pain* (pp. 61–88). Washington, DC: American Psychological Association.
- Böhner, P. (1943). The notitia intuitiva of nonexistents according to William Ockham. *Traditio*, 1, 223–275.
- Boik, R. J. (1979). Interactions, partial interactions, and interaction contrasts in the analysis of variance. *Psychological Bulletin*, 86, 1084–1089.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equations models. *Structural Equation Modeling*, 5, 267–293.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Bonferroni correction: <http://www.cmh.edu/stats/ask/bonferroni.asp>
- Bonferroni correction/adjustment: <http://home.clara.net/sisa/bonhlp.htm>
- Boodoo, G. (1993). Performance-based assessment or multiple choice? *Educational Horizons*, 72(1), 50–56.
- Boole's inequality and Bonferroni's inequalities description: http://www.absoluteastronomy.com/encyclopedia/b/bo/booles_inequality.htm
- Booth, J., Hobert, J., & Jank, W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1, 333–349.
- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer Verlag.
- Borgatta, E. F. (1955). An error ratio for scalogram analysis. *Public Opinion Quarterly*, 19, 96–100.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variable estimation when the correlation between the instruments and endogenous explanatory variables is weak. *Journal of the American Statistical Association*, 90, 443–450.
- Bover, C. B. (1991). *A history of mathematics* (2nd ed.). New York: Wiley.
- Bowen, R. W. (1992). *Graph it! How to make, read, and interpret graphs*. Upper Saddle River, NJ: Prentice Hall.
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572–574.
- Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The Kernel Approach with S-Plus illustrations*. New York: Oxford University Press.
- Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley.
- Boys, A., Marsden, J., Stillwell, G., Hatchings, K., Griffiths, P., & Farrell, M. (2003). Minimizing respondent attrition in longitudinal research: Practical implications from a cohort study of adolescent drinking. *Journal of Adolescence*, 26, 363–373.
- Bracewell, R. (1999). *The Fourier transform and its applications* (3rd ed.). New York: McGraw-Hill.
- Bracken, B. A. (1986). Incidence of basic concepts in five commonly used American tests of intelligence. *School Psychology International*, 7, 1–10.
- Bracken, B. A. (1992). *Multidimensional self concept scale*. Austin, TX: PROED.

- Bracken, B. A. (1996). Clinical applications of a multidimensional, context-dependent model of self-concept. In B. A. Bracken (Ed.), *Handbook of self concept: Developmental, social, and clinical considerations* (pp. 463–505). New York: Wiley.
- Bracken, B. A. (1996). *Handbook of self-concept: Developmental, social, and clinical considerations*. New York: Wiley.
- Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised*. San Antonio, TX: Psychological Corporation.
- Bracken, B. A., Barona, A., Bauermeister, J. J., Howell, K. K., Poggioli, L., & Puente, A. (1990). Multinational validation of the Bracken Basic Concept Scale. *Journal of School Psychology, 28*, 325–341.
- Bracken, B. A., & Boatwright, B. S. (2005). *Clinical assessment of attention deficit—adult*. Lutz, FL: Psychological Assessment Resources.
- Bracken, B. A., & Howell, K. K. (2005). *Clinical assessment of depression*. Lutz, FL: Psychological Assessment Resources.
- Bracken, B. A., Howell, K. K., Harrison, T. E., Stanford, L. D., & Zahn, B. H. (1991). Ipsative subtest pattern stability of the Bracken Basic Concept Scale and the Kaufman Assessment Battery for Children in a preschool sample. *School Psychology Review, 20*, 309–324.
- Bracken, B. A., & Keith, L. K. (2004). *Clinical assessment of behavior*. Lutz, FL: Psychological Assessment Resources.
- Bracken, B. A., & Naglieri, J. A. (2003). Assessing diverse populations with nonverbal tests of general intelligence. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (2nd ed., pp. 243–274). New York: Guilford.
- Bracy, O. L., Oakes, A. L., Cooper, R. S., Watkins, D., Watkins, M., Brown, D. E., & Jewell, C. (1999). The effects of cognitive rehabilitation therapy techniques for enhancing the cognitive/intellectual functioning of seventh and eighth grade children. *International Journal of Cognitive Technology, 4*, 19–27.
- Braddock, C. H., III, Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *Journal of the American Medical Association, 282*(24), 2313–2320.
- Braden, J. P., & Niebling, B. C. (2005). Using the joint test standards to evaluate the validity evidence for intelligence tests. In D. F. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed.). New York: Guilford.
- Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi square test of independence in $r \times c$ tables that have small expected frequencies. *Psychological Bulletin, 86*, 1290–1297.
- Bratley, P., Fox, B. L., & Schrage, L. E. (1987). *A guide to simulation* (2nd ed.). New York: Springer-Verlag.
- Brazelton, T. B., & Nugent, J. K. (1995). *Neonatal Behavioral Assessment Scale* (3rd ed.). London: MacKeith Press.
- Bregler, C., & Omohundro, M. (1994). Surface learning with applications to lipreading. In J. D. Cowan, G. Tesauro, & J. Alspecter (Eds.), *Advances in neural information processing systems*. San Mateo, CA: Morgan Kaufman.
- Breiman, L. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman & Hall/CRC.
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
- Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. New York: Springer.
- Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics, 30*, 89–99.
- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research, Volume I. The analysis of case-control studies*. Lyon, France: International Agency for Research on Cancer.
- Brewer, K. R. W. (2002). *Combined survey sampling inference: Weighing Basu's elephants*. London: Oxford University Press/Arnold.
- Breyfogle, F. W., III, Cupello, J. M., & Meadows, B. (2001). *Managing Six Sigma: A practical guide to understanding, assessing, and implementing the strategy that yields bottom-line success*. New York: Wiley.
- Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge, UK: Cambridge University Press.
- Broadbooks, W. J., & Elmore, P. B. (1987). A Monte Carlo study of the sampling distribution of the congruence coefficient. *Educational and Psychological Measurement, 47*, 1–11.
- Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective inefficiency. *Psychometrika, 14*, 169–182.
- Brookes, M. (2004). *Extreme measures: The dark visions and bright ideas of Francis Galton*. London: Bloomsbury.
- Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities, 38*, 255–270.
- Brown, D., & Crace, R. K. (2002). *Facilitator's guide to the Life Values Inventory*. Williamsburg, VA: Applied Psychology Resources.
- Brown, S. R. (1971). The forced-free distinction in Q-technique. *Journal of Educational Measurement, 8*, 283–287.
- Brown, S. R. (1986). Q technique and method. In W. D. Berry & M. S. Lewis-Beck (Eds.), *New tools for social scientists*. Beverly Hills, CA: Sage.

- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21, 230–258.
- Bruno, J. E. (1986). Assessing the knowledge base of students with admissible probability measurement (APM): A microcomputer based information theoretic approach to testing. *Measurement and Evaluation in Counseling and Development*, 19, 116–130.
- Bruno, J. E. (1989). Monitoring the academic progress of low achieving students: A analysis of right-wrong (R-W) versus Information Referenced (MCW-APM) formative and summative evaluation procedures. *Journal of Research and Development in Education*, 23, 51–61.
- Bruno, J. E. (1996). Time perceptions and time allocation preferences among adolescent boys and girls. *Journal of Adolescence*, 31(121), 109–126.
- Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational & Psychological Measurement*, 55, 959–966.
- Bryant, P. G., & Smith, M. A. (1995). *Practical data analysis: Case studies in business statistics*. Homewood, IL: Richard D. Irwin.
- Buj, V. (1981). Average IQ values in various European countries. *Personality and Individual Differences*, 2, 169–170.
- Bünin, H. (2001). Kolmogorov-Smirnov and Cramer von Mises type two-sample tests with various weights. *Communications in Statistics—Theory and Methods*, 30, 847–866.
- Burckhardt, C. S., & Jones, K. D. (2003). Adult measures of pain: The McGill Pain Questionnaire (MPQ), Rheumatoid Arthritis Pain Scale (RAPS), Short-Form McGill Pain Questionnaire (SF-MPQ), Verbal Descriptive Scale (VDS), Visual Analog Scale (VAS), and West Haven-Yale Multidisciplinary Pain Inventory (WHYMPI). *Arthritis & Rheumatism (Arthritis Care & Research)*, 49, 96–104.
- Burden, R. L., & Faires, J. D. (2004). *Numerical analysis* (8th ed.). Pacific Grove, CA: Brooks/Cole.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.
- Burns, R. C. (1982). *Self-growth in families: Kinetic Family Drawings (KFD) research and application*. New York: Bruner/Mazel.
- Buros, O. K. (Ed.). (1938). *The nineteen thirty eight mental measurements yearbook*. New Brunswick, NJ: Rutgers University Press.
- Buros Center for Testing: <http://www.unl.edu/buros/>
- Buros Institute for Mental Measurements: <http://www.unl.edu/buros/>
- Buros Institute Test Reviews Online: <http://www.unl.edu/buros/>
- Burt, C. (1948). Factor analysis and canonical correlations. *British Journal of Psychology, Statistical Section*, 1, 95–106.
- Burton, D. M. (1997). *The history of mathematics* (3rd ed.). New York: McGraw-Hill.
- Busemeyer, J., & Jones, L. R. (1983). Analysis of multiplicative causal rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562.
- Buthcher, J. N. (2004). *A beginner's guide to the MMPI-2*. Washington, DC: American Psychological Association.
- Buzas, J., Tosteson, T., & Stefanski, L. (2003). *Measurement error*. Institute of Statistics Mimeo Series No. 2544. Raleigh: North Carolina State University.
- Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation*. Washington, DC: American Psychological Association.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Cajori, F. (1919). *A history of mathematics* (2nd ed.). New York: Macmillan.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Campbell, J. (1998). Test reviews. *Journal of Psycho-educational Assessment*, 16, 334–338.
- Campbell, J., Bell, S., & Keith, L. (2001). Concurrent validity of the Peabody Picture Vocabulary Test-Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment*, 8(1), 85–94.
- Campbell, J. M., Ferguson, J. E., Herzinger, C. V., Jackson, J. N., & Marino, C. A. (2004). Combined descriptive and explanatory information improves peers' perceptions of autism. *Research in Developmental Disabilities*, 25(4), 321–339.
- The Campbell Collaboration: <http://www.campbellcollaboration.org/>
- Cantrell, C. (1999). Variable deletion. In B. Thompson (Ed.), *Advances in social science methodology*, Vol. 5 (pp. 321–333). Greenwich, CT: JAI Press.
- Capraro, M. M., & Capraro, R. M. (2001). Bigger is not better: Seeking parsimony in canonical correlation analysis via variable deletion strategies. *Multiple Linear Regression Viewpoints*, 27(2), 24–33.
- Capraro, R. M., & Capraro, M. M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, 62, 771–782.
- Career Assessment Inventory—The Enhanced Version: http://www.pearsonassessments.com/tests/cai_e.htm

- Carl Friedrich Gauss article: http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss
- Carlo Emilio Bonferroni biography and readings: <http://www.aghmed.fsnet.co.uk/bonf/bonf.html>
- Carney, R. N., & Morse, D. T. (2005). [Reviews of the Stanford Achievement Test, 10th Edition]. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 969–975). Lincoln, NE: Buros Institute of Mental Measurements.
- Carr, D., Wegman, E., & Luo, Q. (1997). *ExplorN: Design considerations past and present*. Technical Report 137, Center for Computational Statistics, George Mason University, Fairfax, VA.
- Carroll, B. J. (1981). The Carroll Rating Scale for Depression: I. Development, reliability, and validation. *British Journal of Psychiatry*, 138, 194–200.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Carroll, J. D., & Green, P. E. (1997). *Mathematical tools for applied multivariate analysis*. San Diego, CA: Academic Press.
- Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman & Hall/CRC.
- Carroll, R. T. (2003). *The skeptic's dictionary: A collection of strange beliefs, amusing deceptions, and dangerous delusions*. Chichester, UK: Wiley. Retrieved August 7, 2005, from <http://www.skepdic.com/placebo.html>
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.
- Castagano, K. S. (2001). Special Olympics unified sports: Changes in male athletes during a basketball season. *Adapted Physical Activity Quarterly*, 18(2), 193–206.
- Catoni, O. (1991). Sharp large deviations estimates for simulated annealing algorithms. *Annales de l'institut Henri Poincaré, Probabilités et Statistiques*, 27(3), 291–383.
- Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, 11, 63–65.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton-Mifflin.
- Cattell, R. B. (1980). They talk of some strict testing of us—pish. *Behavioral and Brain Sciences*, 3, 336–337.
- Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: North-Holland.
- Cattell-Horn-Carroll Human Cognitive Abilities Project: www.iapsych.com/chchca.htm
- Cawsey, T. F., Reed, P. L., & Reddon, J. R. (1982). Human needs and job satisfaction: A multidimensional approach. *Human Relations*, 35, 703–715.
- CBM testing materials page: <http://aimsweb.com>
- Cecil Reynolds and BASC-2: <http://www.agsnet.com/psych/oct04a.asp>
- Center for Creative Learning. (n.d.). Creativity assessment test number 72. Retrieved August 30, 2005, from <http://www.creativelearning.com/Assess/test72.htm>
- Centers for Disease Control and Prevention. (2003, September 5). Cigarette smoking-attributable morbidity—United States, 2000. *MMWR*, 52(35), 842–844. Available from <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5235a4.htm>
- Chall, M., & Owen, T. M. (1995). Documenting the world's sociological literature: Sociological Abstracts. *Publishing Research Quarterly*, 11(3), 83–95.
- Chance magazine: <http://www.amstat.org/publications/chance/>
- Chance Web site devoted to the teaching of courses dealing with chance: <http://www.dartmouth.edu/~chance/>
- Chang, H., & Ying, Z. (1999). a-Stratified computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.
- Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67, 387–398.
- Charmaz, K. (2000). Grounded theory: Constructivist and objectivist methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509–535). Thousand Oaks, CA: Sage.
- Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). Toronto, Canada: Chapman & Hall.
- Chaya, C., Perez-Hugalde, C., Judez, L., Wee, C. S., & Guinard, J. X. (2003). Use of the STATIS method to analyze time-intensity profiling data. *Food Quality and Preference*, 15, 3–12.
- Chessel, D., Dufour, A.-B., & Thioulouse, J. (2004). The ade4 package-I: One-table methods. *R News*, 4, 5–10.
- Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Sarasota, FL: Authors Cooperative.
- Chi-square calculator: <http://faculty.vassar.edu/lowry/csfit.html>
- Choi, B. C. K., & Pak, A. W. P. (1998). Bias, overview. In P. Armitage & P. Colton (Eds.), *Encyclopedia of biostatistics: Vol. 1* (pp. 331–338). Chichester, UK: Wiley.
- Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. *Preventing Chronic Diseases*, 2, 1–20. Retrieved August 6, 2005, from http://www.cdc.gov/pcd/issues/2005/jan/04_0050.htm
- Chow, S. L. (1991). Some reservations about statistical power. *American Psychologist*, 46, 1088–1089.
- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.
- Christensen, A. L. (1974). *Luria's neuropsychological investigation*. Copenhagen: Munksgaard.
- Citro, C. F., Ilgen, D. R., & Marrett, C. B. (Eds.). (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academies Press.

- Clark, R. (2004). Interethnic group and intra-ethnic group racism: Perceptions and coping in Black university students. *Journal of Black Psychology*, 30, 506–526.
- Class interval discussion: Scale and impression: <http://www.shodor.org/interactivate/discussions/sd2.html>
- Classification and regression tree downloads: <http://cran.au.r-project.org/src/contrib/Descriptions/tree.html>
- Clausen, S. E. (1998). *Applied correspondence analysis*. Thousand Oaks, CA: Sage.
- Clayton, M. J. (1997). Delphi: A technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology*, 17(4), 373–386.
- Clerical-type work: www.bls.gov/oco/ocos130.htm#nature
- Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.
- Cleveland, W. S., & McGill, M. E. (Eds.). (1988). *Dynamic graphics for statistics*. Pacific Grove, CA: Wadsworth.
- Cleveland, W. S., & McGill, R. (1984). The many faces of the scatterplot. *Journal of the American Statistical Association*, 79, 807–822.
- Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18, 115–126.
- Clogg, C. C., & Goodman, L. A. (1984). Latent structure-analysis of a set of multidimensional contingency-tables. *Journal of the American Statistical Association*, 79(388), 762–771.
- Clustering algorithm tutorial: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html
- Cochran, E. L. (1994). Basic versus applied research in cognitive science: A view from industry. *Ecological Psychology*, 6, 131–135.
- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256–266.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- The Cochrane Collaboration: The reliable source of evidence in health care: <http://www.cochrane.org/index0.htm>
- CogAT recent research links: www.cogat.com, <http://faculty.education.uiowa.edu/dlohman>
- Cohen, B. H., & Lea, R. B. (2004). *Essentials of statistics for the social and behavioral sciences*. Hoboken, NJ: Wiley.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6), 426–443.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858–866.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., Aiken, L., & West, S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cole, D. J., Ryan, C. W., & Kick, F. (1995). *Portfolios across the curriculum and beyond*. Thousand Oaks, CA: Corwin.
- Collier, R. O., Jr., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios, in repeated measures designs. *Psychometrika*, 32, 339–353.
- Colom, R., & García-López, O. (2003). Secular gains in fluid intelligence: Evidence from the Culture Fair Intelligence Test. *Journal of Biosocial Science*, 35, 33–39.
- Comprehensive Meta-Analysis: A computer program for research synthesis: <http://www.meta-analysis.com/>
- Comrey, A. L. (1970). *Manual for the Comrey Personality Scales*. San Diego, CA: Educational and Industrial Testing Service.
- Comrey, A. L. (1976). Mental testing and the logic of measurement. In W. L. Barnette (Ed.), *Readings in psychological tests and measurement*. Baltimore: Williams & Wilkins.
- Comrey, A. L. (1995). *Manual and handbook of interpretations for the Comrey Personality Scales*. San Diego, CA: EDITS Publishers.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Comrey, A. L., & Lee, H. B. (1995). *Elementary statistics: A problem solving approach* (3rd ed.). Dubuque, IA: Kendall-Hunt.
- Conditional probability: http://en.wikipedia.org/wiki/Conditional_probability
- Confidence intervals, Rice Virtual Lab in Statistics page: http://davidmlane.com/hyperstat/confidence_intervals.html
- Connotative meaning information: <http://www.writing.ws/reference/history.htm>
- Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Consulting Psychologists Press, Inc.: www.cpp-db.com
- Cook, D. (1997). Calibrate your eyes to recognize high-dimensional shapes from their low-dimensional

- projections [Electronic version]. *Journal of Statistical Software*, 2(6).
- Cook, R. D. (2000, June). *Using Arc for dimension reduction and graphical exploration in regression*. Retrieved from <http://www.stat.umn.edu/arc/InvReg/DimRed.pdf>
- Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Society*, 100, 410–428.
- Cook, R., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage.
- Coquet, R., Troxler, L., & Wipff, G. (1996). The STATIS method: Characterization of conformational states of flexible molecules from molecular dynamics simulation in solution. *Journal of Molecular Graphics*, 14, 206–212.
- Corno, L., Cronbach, L. J., Lohman, D. F., Kupermintz, H., Mandinach, E. B., Porteus, A., et al. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Hillsdale, NJ: Erlbaum.
- Correlation coefficient page: <http://mathworld.wolfram.com/CorrelationCoefficient.html>
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor-Inventory Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Council for Exceptional Children: <http://www.cec.sped.org>
- Cowie, A. P. (1989). *Oxford advanced learner's dictionary*. Oxford, UK: Oxford University Press.
- Cowton, C. J. (1998). The use of secondary data in business ethics research. *Journal of Business Ethics*, 17, 423–434.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society B*, 33, 187–220.
- Cox, D., & Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, B*, 30, 248–275.
- Cozby, P. C. (1985). *Methods in behavioral research* (3rd ed.). Palo Alto, CA: Mayfield.
- Crace, R. K., & Brown, D. (2002). *Understanding your values*. Williamsburg, VA: Applied Psychology Resources.
- Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. *Gifted Child Quarterly*, 49, 283–291.
- Creative Research Systems. (2003). *The survey system*. Retrieved August 23, 2005, from <http://www.survey-system.com/sdesign.htm>
- Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.
- Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice*, 39(3), 124–130.
- The CRISP-DM consortium: <http://www.crisp-dm.org/>
- Cristianini, N., & Shawe-Taylor, J. (2001). *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press.
- Crites, J. O. (1978). *Theory and research handbook for the Career Maturity Inventory* (2nd ed.). Monterey, CA: CTB/McGraw-Hill.
- Crites, J. O., & Savickas, M. L. (1996). Revision of the Career Maturity Inventory. *Journal of Career Assessment*, 4, 131–138.
- Crites, S. L., Jr., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, 20(6), 619–634.
- Critical *r* values table: www.ecowin.org/aulas/resources/stats/correlation.htm
- Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3), 5–11.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1989). Lee J. Cronbach. In G. Lindzey (Ed.), *A history of psychology in autobiography, Vol. 8* (pp. 62–93). Stanford, CA: Stanford University Press.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Crowder, M. J., & Hand, D. J. (1990). *Analysis of repeated measures*. London: Chapman & Hall.
- Cuddeback, G., Wilson, E., Orme, J. G., & Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30, 19–33.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions.

- Educational and Psychological Measurement*, 61, 532–575.
- Cummings, O. W. (1981). *Validation of a diagnostic interpretation technique for the Iowa Test of Basic Skills: Final report to the National Institute of Education*. Grant Wood Area Education Agency.
- Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, 21, 287–290.
- Cureton, E. E., Cronbach, L. J., Meehl, P. E., Ebel, R. L., Ward, A. W., & Stoker, H. W. (1996). Validity. In A. W. Ward, H. W. Stoker, & M. Murray-Ward (Eds.), *Educational measurement: Origins, theories, and explanations: Vol. 1. Basic concepts and theories*. Lanham, MD: University Press of America.
- Curtice, J., & Sparrow, N. (1997). How accurate are traditional quota opinion polls? *Journal of the Market Research Society*, 39(3), 433–438.
- Curvilinear bivariate regression: <http://core.ecu.edu/psyc/wuenschk/MV/multReg/Curvi.doc>
- Curvilinear regression: http://www.vias.org/tmdatanaleng/cc_regress_curvilinear.html
- D'Agostino, R. B. (1986). Tests for the normal distribution. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 367–419). New York: Marcel Dekker.
- D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician*, 44, 316–321.
- D'Agostino, R. B., & Stephens, M. A. (Eds.). (1986). *Goodness-of-fit-techniques*. New York: Marcel Dekker.
- Dagnelie, P. (1968). A propos de l'emploi du test de Kolmogorov-Smirnov comme test de normalité. *Biometrie-Praximetrie*, 9(1), 3–13.
- DAMBE Windows95/98/NT executables: <http://aix1.uottawa.ca/~xxia/software/software.htm>
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Darlington, R. B., & Hayes, A. F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. *Psychological Methods*, 4, 496–515.
- Darwin: <http://pages.britishlibrary.net/charles.darwin/>
- Das Gupta, S. (1960). Point biserial correlation and its generalization. *Psychometrika*, 25, 393–408.
- Data compression information: <http://datacompression.info>
- Data compression principles and practices: <http://www.data-compression.com>
- Data mining information clearinghouse: <http://www.kdnuggets.com/>
- Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). Perseveration of checking thoughts and mood-as-input hypothesis. *Journal of Behavior Therapy & Experimental Psychiatry*, 34, 141–160.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.
- Dawson, R. J. M. (1995). The “unusual episode” data revisited. *Journal of Statistics Education*, 3. Available: <http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html>
- Day, S. J. (1947). *Intuitive cognition: A key to the significance of the late Scholastics*. St. Bonaventure, NY: Franciscan Institute.
- De Bildt, A., Kraijer, D., Sytema, S., & Minderaa, R. (2005). The psychometric properties of the Vineland Adaptive Behavior Scales in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders*, 35, 53–62.
- de Castro, J. M., & Brewer, E. M. (1991). The amount eaten in meals by humans is a power function of the number of people present. *Physiology & Behavior*, 51, 121–125.
- De Gagne, T. A., Mikail, S. F., & D'Eon, J. L. (1995). Confirmatory factor analysis of a 4-factor model of chronic pain evaluation. *Pain*, 60, 195–202.
- De Gruttola, V., & Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50, 1003–1014.
- de la Torre, J., & Douglas, J. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333–353.
- De Leeuw, J. (1983). On the prehistory of correspondence analysis. *Statistica Neerlandica*, 37, 161–164.
- Dear, R. E. (1959). *A principal-components missing data method for multiple regression models* (Technical Report SP-86). Santa Monica, CA: Systems Development Corporation.
- Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence*, 23, 105–132.
- Degenne, A. (1972). *Techniques ordinales en analyse des données*. Paris: Hachette.
- Dekking, M., Lévy Véhel, J., Lutton, E., & Tricot, C. (Eds.). (1999). *Fractals: Theory and applications in engineering*. London: Springer.
- Dembo, A., & Karlin, S. (1992). Poisson approximations for r -scan processes. *Annals of Applied Probability*, 2(2), 329–357.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Department of Health and Human Services—Centers for Disease Control and Prevention: www.cdc.gov
- Descriptive Statistics: <http://www.physics.csbsju.edu/stats/descriptive2.html> (application that computes the mean once data are entered)
- Detterman, D. K., & Daniel, M. H. (1989). Correlates of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, 13, 349–359.

- DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2005). *Intro stats* (2nd ed.). Boston: Addison-Wesley.
- Dey, I. (1999). *Grounding grounded theory*. San Diego, CA: Academic Press.
- Differential Aptitude Tests for Personnel and Career Assessment: <http://www.pantesting.com/products/PsychCorp/DAT.asp>
- Differential Aptitudes Tests: www.psychcorpcenter.com
- Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*. London: Hodder Arnold.
- Diggle, P. J., Heagerty, P. J., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford, UK: Clarendon.
- Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49–93.
- Dijksterhuis, G. B., & Gower, J. C. (1991/2). The interpretation of generalized procrustes analysis and allied methods. *Food Quality and Preference*, 3, 67–87.
- Ding, C. S. (2001). Profile analysis: Multidimensional scaling approach. *Practical Assessment, Research & Evaluation*, 7(16). Retrieved October 1, 2005, from <http://PAREonline.net/getvn.asp?v=7&n=16>
- Disabilities Education Improvement Act (IDEA) of 2004 resources: <http://www.ed.gov/offices/OSERS/IDEA/>
- Disease rates comparison: <http://bmj.bmjournals.com/epidem/epid.3.html>
- Dixon, W. J. (1951). Ratios involving extreme values. *Annals of Mathematical Statistics*, 22(1), 68–78.
- Dixon test for outliers has been implemented in the R project, a free software environment for statistical computing and graphics: <http://finzi.psych.upenn.edu/R/library/outliers/html/dixon.test.html>
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434.
- Doll, R., Peto, R., Wheatley, K., Gray, R., & Sutherland, I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal*, 309(6959), 901–911.
- Donnay, D., Morris, M., Schaubhut, N., & Thompson, R. (2005). *Strong Interest Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22, 33–51.
- Dotson, L. E., & Summers, G. F. (1970). Elaboration of Guttman scaling techniques. In G. F. Summers (Ed.), *Attitude measurement* (pp. 203–213). Chicago: Rand McNally.
- Dougherty, D. M., Marsh, D. M., & Mathias, C. W. (2002). Immediate and Delayed Memory Tasks: A computerized measure of memory, attention, and impulsivity. *Behavioral Research Methods, Instruments, and Computers*, 34, 391–398.
- Dougherty, D. M., Mathias, C. W., & Marsh, D. M. (2003). Laboratory measures of impulsivity. In E. F. Coccaro (Ed.), *Aggression: Psychiatric assessment and treatment* (Medical Psychiatric Series No. 22, pp. 247–265). New York: Marcel Dekker.
- Dougherty, D. M., Mathias, C. W., Marsh, D. M., & Jagar, A. A. (2005). Laboratory behavioral measures of impulsivity. *Behavior Research Methods*, 37, 82–90.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Dr. Levant: <http://www.DrRonaldLevant.com>
- Drawing tree diagrams: <http://www-stat.stanford.edu/~susan/surprise/ProbabilityTree.html>
- Droitcour, J. A., Larson, E. M., & Scheuren, F. J. (2001). The three card method: Estimating sensitive survey items—with permanent anonymity of response. *Proceedings of the American Statistical Association, Statistical Computing Section*, CD-ROM. Retrieved from <http://www.amstat.org/sections/srms/Proceedings/y2001/Proceed/00582.pdf>
- Drug-Free Workplace Act of 1988, 41 U.S.C. Sec. 701 et seq.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance-based assessments. *Applied Measurement in Education*, 4(4), 289–303.
- Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.
- Duncan, O. D., & Hodge, R. W. (1963). Education and occupational mobility: A regression analysis. *American Journal of Sociology*, 68, 629–644.
- Dunlap, W. P., Burke, M. J., & Smith Crowe, K. (2003). Accurate tests of statistical significance for r sub(WG) and average deviation interrater agreement indexes. *Journal of Applied Psychology*, 88(2), 356–362.
- Dunn, O. L. (1961). Multiple comparisons of means. *Journal of the American Statistical Association*, 56, 52–64.
- Dunn, O. L., & Massey, F. J., Jr. (1965). Estimation of multiple contrasts using t -distributions. *Journal of the American Statistical Association*, 60, 573–583.
- Dunn-Rankin, P., Knezek, G. A., Wallace, S., & Zhang, S. (2004). *Scaling methods* (2nd ed.). Mahwah, NJ: Erlbaum.
- Dunnington, G. W. (2004). *Carl Friedrich Gauss: Titan of science*. Washington, DC: Mathematical Association of America.
- Dyer, F. N. (1973). The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. *Memory & Cognition*, 1(2), 106–120.
- E. Paul Torrance: <http://www.coe.uga.edu/torrance/>
- Eaves, R. C., McLaughlin, P. J., & Foster, G. G. (1979). Some simple statistical procedures for classroom use. *Diagnostic*, 4, 3–12.

- Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement*, 16, 269–282.
- Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, 66, 485–487.
- EdITS/Educational and Industrial Testing Service: <http://www.edits.net>
- Educational Testing Service: <http://www.ets.org>
- Edwards, A. L. (1957). *Techniques of attitude scale construction* (pp. 172–199). New York: Appleton-Century-Crofts.
- Edwards, A. L. (1959). *Manual: Edwards Personal Preference Schedule*. Washington, DC: The Psychological Corporation.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Elementary and Secondary Education Act, 20 U.S.C. 6311(b)(3)(C)(ii) (2002).
- Elliott, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review*, 26, 224–233.
- Elliott, S. N., & Gresham, F. M. (1991). *Social skills intervention guide: Practical strategies for social skills training*. Circle Pines, MN: American Guidance Service.
- Elsholtz, J. S. (1654). *Anthropometria*. Padua: M. Cadorini.
- Embretson, S. E. (1992). *Technical manual for the Spatial Learning Ability Test* (Tech. Rep. No. 9201). Lawrence: University of Kansas, Department of Psychology.
- Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective*. New York: Plenum.
- Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32, 277–294.
- Embretson, S. E. (1996). Cognitive design systems and the successful performer: A study on spatial ability. *Journal of Educational Measurement*, 33, 29–39.
- Embretson, S. E. (1997). The factorial validity of scores from a cognitively designed test: The Spatial Learning Ability Test. *Educational and Psychological Measurement*, 57, 99–107.
- Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Emerson, J. D., & Hoaglin, D. C. (1983). Stem-and-leaf displays. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 7–32). New York: Wiley.
- Enders, W. (2004). *Applied econometric time series* (2nd ed.). Hoboken, NJ: Wiley.
- Engelhard, G., Jr. (1994). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 73–99). Norwood, NJ: Ablex.
- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistiques Appliquées*, 26, 29–37.
- Escofier, B., & Pagès, J. (1990). *Analyses factorielles simples et multiples: Objectifs, méthodes, interprétation*. Paris: Dunod.
- Escofier, B., & Pagès, J. (1994). Multiple factor analysis. *Computational Statistics & Data Analysis*, 18, 121–140.
- Escofier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples*. Paris: Dunod.
- Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29, 751–760.
- Escoufier, Y. (1980). L'analyse conjointe de plusieurs matrices de données. In M. Jolivet (Ed.), *Biométrie et temps* (pp. 59–76). Paris: Société Française de Biométrie.
- Etheridge, E. W. (1992). *Sentinel for health: A history of the Centers for Disease Control*. Berkeley: University of California Press.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing* (2nd ed.). New York: Dekker.
- Evans, J. D. (1985). *Invitation to psychological research*. New York: Holt, Rinehart, & Winston.
- Evans, M. G. (1985). A Monte-Carlo study of correlated error in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes*, 36, 305–323.
- Evans, M. G. (1991). The problem of analyzing multiplicative composites: Interactions revisited. *American Psychologist*, 46, 6–15.
- Evans, M. G. (1999). On the asymmetry of g. *Psychological Reports*, 85, 1059–1069.
- Evans, M. G. (2002). The implications of the asymmetry of g for predictive validity. In W. Auer-Rizzi, E. Szabo, & C. Innreiter-Moser (Eds.), *Management in einer Welt der Globalisierung und Diversität: Europäische und nordamerikanische Sichtweisen* (pp. 433–441). Stuttgart, Germany: Schaeffer-Poeschel Verlag.
- Evans, M. J., & Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. New York: Freeman.
- Everitt, B. (1999). *Chance rules: An informal guide to probability, risk, and statistics*. New York: Springer.
- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London: Hodder Arnold.
- Exner, J. (2003). *The Rorschach: A comprehensive system* (4th ed.). New York: Wiley.
- Exploratory factor analysis primer: http://www.upa.pdx.edu/IOA/newsom/semclass/ho_efa.doc
- Exploratory Software for Confidence Intervals: <http://www.latrobe.edu.au/psy/esci/>

- Exploring Careers: The ASVAB career exploration guide.* (2005). DD Form 1304-5WB, July 2005. U.S. Government Printing Office.
- ExplorN software for touring through high-dimensional data using parallel coordinate plots: <ftp://www.galaxy.gmu.edu/pub/software/> (It is now replaced by its commercial evolution, CrystalVision)
- Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In C. T. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 17–40). Thousand Oaks, CA: Sage.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Factor analysis article: http://en.wikipedia.org/wiki/Factor_analysis
- Factor analysis software sampling: <http://quantrm2.psychology.okstate.edu/faculty/jgrice/factorscores/>
- Faden, R., & Beauchamp, T. (1986). *A history and theory of informed consent*. New York: Oxford University Press.
- Fagan, J. F., & Detterman, D. K. (1992). The Fagan Test of Infant Intelligence: A technical summary. *Journal of Applied Developmental Psychology*, 13, 173–193.
- Fahome, G. (2002). Twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods*, 1(2), 248–268.
- Fahome, G., & Sawilowsky, S. (2000, April). *Twenty nonparametric statistics*. Paper presented at the annual meeting of the American Educational Research Association, SIG/Educational Statisticians, New Orleans, LA.
- FairTest: <http://www.fairtest.org>
- False negative: http://en.wikipedia.org/wiki/Type_II_error
- Family Education Rights and Privacy Act (29 U.S.C.A. § 1232g).
- Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, 63, 915–930.
- Fay, B. R. (2006). The effect Type I error and power of various methods of resolving ties for six distribution-free tests of location. *Journal of Modern Applied Statistical Methods*, 5(1) 50–67.
- Feder, J. (1988). *Fractals*. New York: Plenum.
- Federal regulations on the protection of human participants: <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm>
- Feinberg, M., Carroll, B. J., & Smouse, P. E. (1981). The Carroll Rating Scale for Depression: III. Comparison with other rating instruments. *British Journal of Psychiatry*, 138, 205–209.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education; Macmillan.
- Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education*, 11, 159–177.
- Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11, 93–103.
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183–1210.
- Fewer, L. J., Howard, C., & Shepherd, R. (1997). Public concerns in the United Kingdom about general and specific applications of genetic engineering: Risk, benefit, and ethics. *Science, Technology, & Human Values*, 22, 98–124.
- Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
- Field, S., & Hoffman, A. (1994). Development of a model for self-determination. *Career Development for Exceptional Individuals*, 17, 159–169.
- Field, S., Hoffman, A., & Sawilowsky, S. (1995). *Self-Determination Knowledge Scale, Form A and Form B*. Austin, TX: Pro-Ed.
- Fienberg, S. E. (1994). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Figueroa, R. (1982). SOMPA and the psychological testing of Hispanic children. *Metas*, 2, 1–6.
- Figueroa, R. A., & Sassenrath, J. M. (1989). A longitudinal study of the predictive validity of the System of Multicultural Pluralistic Assessment (SOMPA) (ERIC Document Reproduction Service No. EJ391800). *Psychology in the Schools*, 26, 5–19.
- Fill, J. (1988). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 8, 131–162.
- Finch, A. J., & Belter, R. W. (1993). Projective techniques. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child and adolescent assessment* (pp. 224–238). Boston: Allyn & Bacon.
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, 31, 439–485.
- Finch, S. M. (1960). *Fundamentals of child psychiatry*. New York: Norton.
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the journal of applied psychology: Little evidence of reform. *Educational and Psychological Measurement*, 61, 181–210.

- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV® Axis I disorders (SCID-I), clinician version, user's guide*. Washington, DC: American Psychiatric Press.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fischman, M. (2000). Informed consent. In B. Sales & F. Susan (Eds.), *Ethics in research with human participants* (pp. 35–48). Washington, DC: American Psychological Association.
- Fisher, C. B. (2003). *Decoding the ethics code: A practical guide for psychologists*. Thousand Oaks, CA: Sage.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10, 507–521.
- Fisher, R. A. (1925). *Statistical methods for research workers* (1st–13th eds.). Edinburgh, UK: Oliver & Boyd.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
- Fisher, R. A. (1971). *The design of experiments*. New York: Hafner. (Original work published in 1935)
- Fisher's exact test calculation: <http://www.unc.edu/~preacher/fisher/fisher.htm>
- Fisher's Z description with formulas: <http://davidmlane.com/hyperstat/A50760.html>
- Fisher's z'-Transformation, from *MathWorld*—a Wolfram Web resource, by E. W. Weisstein: <http://mathworld.wolfram.com/Fishersz-Transformation.html>
- Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112, 393–395.
- Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In E. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237–270). Washington, DC: American Council on Education.
- Flanagan, D. P., Genshaft, J. L., & Harrison, P. L. (1997). *Contemporary intellectual assessment: Theories, tests, and issues*. New York: Guilford.
- Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Boston: Allyn & Bacon.
- Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.
- Flury, B., & Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. New York: Chapman & Hall.
- Fourier transform article: http://en.wikipedia.org/wiki/Fourier_transform
- Fourier transformations information by Wolfram Research: <http://mathworld.wolfram.com/FourierTransform.html>
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Frank Gresham home page: <http://www.behavioralinstitute.org/Frank%20Gresham.htm>
- Frank Wilcoxon biographical essay: <http://www.umass.edu/wsp/statistics/tales/wilcoxon.html>
- Frazer, G. H., & Sechrist, S. R. (1994). A comparison of occupational stressors in selected allied health disciplines. *Health Care Supervisor*, 13(1), 53–65.
- Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York: Norton.
- Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine learning: Proceedings of the Thirteenth International Conference* (L. Saitta, ed., pp. 148–156). San Francisco: Morgan Kaufmann.
- Friedman, A. J., Lewak, R., Nichols, D. S., & Webb, J. T. (2001). *Psychological assessment with the MMPI-2*. Mahwah, NJ: Erlbaum.
- Friedman, D., Putnam, L., Ritter, W., Hamberger, M., & Berman, S. (1992). A developmental event-related potential study of picture matching in children, adolescents, and young adults: A replication and extension. *Psychophysiology*, 29, 593–610.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675–701.
- Friedman's Test Applet (allows you to enter data and calculate the test statistic): <http://www.fon.hum.uva.nl/Service/Statistics/Friedman.html>
- Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32, 124–158.
- Friendly, M. (1995). Conceptual and visual models for categorical data. *American Statistician*, 49, 153–160.
- Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8, 373–395.
- Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41, 103–130.
- Friendly, M., & Denis, D. J. (2006). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Retrieved from <http://www.math.yorku.ca/SCS/Gallery/milestone/>
- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, 43, 50–54.
- FTII description and other measures of infant cognitive assessment: <http://ehp.niehs.nih.gov/members/2003/6205/6205.html#comp>
- Fukunaga, K. (1972). *Introduction to statistical pattern recognition*. New York: Academic Press.
- Galambos, J., & Simonelli, I. (1996). *Bonferroni-type inequalities with applications*. New York: Springer.

- Galton, F. (1878). Composite portraits. *J Anthropol Inst Gr Brit & Ireland*, 8, 132.
- Galton: <http://www.mugu.com/galton/index.html>
- Games, P. A. (1977). An improved *t* table for simultaneous control on *g* contrasts. *Journal of the American Statistical Association*, 72, 531–534.
- Games, P. A., Keselman, H. J., & Clinch, J. J. (1979). Tests for homogeneity of variance in factorial designs. *Psychological Bulletin*, 86, 978–984.
- Garson, J. G. (1884). The Frankfort Craniometric Convention, with critical remarks thereon. *J Anthropol Inst Gr Brit & Ireland*, 14, 64–83.
- Gauss, C. F. (1821). *Theoria combinationis observationum erroribus minimis obnoxiae*. Göttingen, Germany: Royal Society of Göttingen.
- Gee, T. (1998). *Individual and joint-level properties of personal project matrices: An exploration of the nature of project spaces*. Unpublished doctoral dissertation, Carleton University, Ottawa, Canada.
- Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Gentle, J. E. (2002). *Elements of computational statistics*. New York: Springer.
- Gentle, J. E., Härdle, W., & Mori, Y. (Eds.). (2004). *Handbook of computational statistics: Concepts and methods*. New York: Springer.
- Geoffroy, J. (1958, 1959). Contribution a la theorie des valeurs extremes. *Publications de l'Institut de Statistique de l'Universite de Paris*, 7, 37–121; 8, 123–184.
- Ghysels, E., & Osborn, D. R. (2001). *The econometric analysis of seasonal time series*. New York: Cambridge University Press.
- Gibbons, J. D. (1993). *Nonparametric measures of association* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07B091). Newbury Park, CA: Sage.
- Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference* (4th ed.). New York: Dekker.
- Gieser, L., & Stein, M. I. (1999). *Evocative images: The Thematic Apperception Test and the art of projection*. Washington, DC: American Psychological Association.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, UK: Wiley.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606.
- Given, L. M., & Leckie, G. J. (2003). "Sweeping" the library: Mapping the social activity space of the public library. *Library & Information Science Research*, 25(4), 365–385.
- Givens, G., & Hoeting, J. (2005). *Computational statistics*. New York: Wiley Interscience.
- Glaser, B. G. (1992). *Emerging versus forcing: Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Glaser, R., & Klaus, D. J. (1962). Proficiency measurements: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in systems development*. New York: Holt, Rinehart & Winston.
- Glass, G. V., & Hopkins, K. D. (1995). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.
- Glatthorn, A. A. (1998). *Writing the winning dissertation: A step-by-step guide*. Thousand Oaks, CA: Corwin Press.
- Glaz, J., & Balakrishnan, N. (1999). *Scan statistics and applications*. Boston: Birkhauser.
- Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, 40(2), 229–252.
- Glonek, G. F. V., & McCullagh, R. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society—Series B*, 57, 533–546.
- Goffin, R. D., & Helmes, E. (Eds.). (2000). *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy*. Boston: Kluwer.
- Gold, E. (1997). *The gambler's fallacy*. Unpublished doctoral dissertation, Carnegie Mellon University.
- Gold, M. S., Byars, J. A., & Frost-Pineda, K. (2004). Occupational exposure and addictions for physicians: Case studies and theoretical implications. *Psychiatric Clinics of North America*, 27(4), 745–753.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.
- Golden, C. J. (1976). The diagnosis of brain damage by the Stroop test. *Journal of Clinical Psychology*, 32, 654–658.
- Golden, C. J., & Freshwater, S. M. (2001). Luria-Nebraska Neuropsychological Battery. In W. I. Dorfman & M. Hersen (Eds.), *Understanding psychological assessment: Perspectives on individual differences*. New York: Kluwer Academic/Plenum.
- Golden, C. J., Freshwater, S. M., & Vayalakkara, J. (2000). The Luria-Nebraska Neuropsychological Battery. In G. Groth-Marnat (Ed.), *Neuropsychological assessment*

- in clinical practice: A guide to test interpretation and integration*. New York: Wiley.
- Golub, G. H., & van Loan, C. F. (1989). *Matrix computations*. Baltimore: Johns Hopkins University Press.
- Goodenough-Harris drawing test: <http://gri.gallaudet.edu/~catraxle/INTELLEC.html#goodenough>
- Goodman, J. S., & Blum, T. C. (1996). Assessing the nonrandom sampling effects of subject attrition in longitudinal research. *Journal of Management*, 22, 627–652.
- Goodman, L. A. (1974). Exploratory latent structure-analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215–231.
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r . *The Journal of Experimental Education*, 74(3), 251–266.
- Gordon, A. D. (1999). *Classification*. London: Chapman & Hall.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology*, 77, 631–635.
- Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology*, 82, 525–538.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (1994). Role of parental motivational practices in children's academic intrinsic motivation and achievement. *Journal of Educational Psychology*, 86, 104–113.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (1998). Role of cognitively stimulating home environment in children's academic intrinsic motivation: A longitudinal study. *Child Development*, 69, 1448–1460.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, 93, 3–13.
- Gottfried, A. W., Gottfried, A. E., Cook, C., & Morris, P. (2005). Educational characteristics of adolescents with gifted academic intrinsic motivation: A longitudinal study from school entry through early adulthood. *Gifted Child Quarterly*, 49, 172–186.
- Gough, H. G. (1952). *The Adjective Check List*. Palo Alto, CA: Consulting Psychologists Press.
- Gough, H. G. (1996). *The California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Gould, S. J. (1996). *The mismeasure of man* (revised & expanded ed.). New York: Norton.
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.
- Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40, 33–51.
- Grace, C., Shores, E., & Charner, K. (Eds.). (1998). *The portfolio book: A step-by-step guide for teachers*. Beltsville, MD: Gryphon House.
- Graduate Record Examinations Web site: <http://www.gre.org>
- Graduate Record Examinations Board. (1997). *GRE 1997–98: Guide to the use of scores*. Princeton, NJ: Educational Testing Service.
- Graffi, S., & Minnes, P. M. (1988). Attitudes of primary school children toward the physical appearance and labels associated with Down syndrome. *American Journal of Mental Retardation*, 93(1), 28–35.
- Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow up data. *Journal of Applied Psychology*, 78, 119–128.
- Gravetter, F. J., & Wallnau, L. B. (2005). *Essentials of statistics for the behavioral sciences*. Belmont, CA: Wadsworth.
- Great Britain. Parliament. (1998). *Report on the loss of the S.S. Titanic: The official government enquiry*. New York: Picador.
- Green, D., & Shapiro, I. (1994). *Pathologies of rational choice theory: A critique of applications in political science*. New Haven, CT: Yale University Press.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.
- Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Service.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.
- Griffiths, W. E., Hill, R. C., & Judge, G. G. (1993). *Learning and practicing econometrics*. New York: Wiley.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Hillsdale, NJ: Erlbaum.
- Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Macmillan.
- Groth-Marnat, G. (2003). *Handbook of psychological assessment*. New York: Wiley.
- Grounded theory resources: http://dmoz.org/Science/Social_Sciences/Methodology/Grounded_Theory/
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Guo, S., & Hussey, D. L. (2004). Nonprobability sampling in social work research: Dilemmas, consequences, and strategies. *Journal of Social Service Research*, 30, 1–18.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Guttman, L. (1955). The determinacy of factor score matrices with applications for five other problems of common factor theory. *British Journal of Statistical Psychology*, 8, 65–82.
- Guttman, L. (1970). The Cornell technique for scale and intensity analysis. In G. F. Summers (Ed.), *Attitude measurement* (pp. 187–203). Chicago: Rand McNally.
- Gynther, M. D., Burkhart, B., & Hovanitz, C. (1979). Do face validity items have more predictive validity than subtle items? *Journal of Consulting and Clinical Psychology*, 47, 295–300.
- Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Annals of Statistics*, 2(5), 911–924.
- Haberman, S. J. (1977). Product models for frequency tables involving indirect observation. *Annals of Statistics*, 5(6), 1124–1147.
- Haertel, E. H. (1999). Validity arguments for high-stake testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5–9.
- Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.
- Haig, B. D. (1995). *Grounded theory as scientific method* (Philosophy of Education Society Yearbook 1995, pp. 281–290). Urbana: University of Illinois Press.
- Haig, B. D. (2003). What is a spurious correlation? *Understanding Statistics*, 2, 125–132.
- Haight, F. A. (1967). *Handbook of the Poisson distribution*. New York: Wiley.
- Haining, R. P. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge, UK: Cambridge University Press.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed., pp. 469–519). Upper Saddle River, NJ: Prentice Hall.
- Hajek, J. (1969). *A course in nonparametric statistics*. San Francisco: Holden-Day.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.
- Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple choice item writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334.
- Hall, J. M., & Rosenthal, R. (1991). Testing for moderator variables in meta analysis: Issues and methods. *Communication Monographs*, 58, 437–448.
- Hall, W. B., & Gough, H. G. (1977). Selecting statistical clerks with the Minnesota Clerical Test. *Journal of Psychology*, 96, 297–301.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Belmont, CA: Duxbury.
- Hamilton, W., & Burns, T. G. (2003). WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence (3rd ed.). *Applied Neuropsychology*, 10(3), 188–190.
- Hand, D. J. (1997). *Construction and assessment of classification rules*. New York: Wiley.
- Hand, D. M. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 159, 445–492.
- Handleman, J. S., & Harris, S. L. (1986). *Educating the developmentally disabled: Meeting the needs of children and families*. San Diego, CA: College-Hill Press.
- Handler, L., & Habenicht, D. (1994). The Kinetic Family Drawing technique: A review of the literature. *Journal of Personality Assessment*, 62, 440–464.
- Hanke, J. E., & Wichern, D. W. (2005). *Business forecasting*. Upper Saddle River, NJ: Pearson.
- Hansen, J. C. (1992). *User's guide for the Strong Interest Inventory*. Stanford, CA: Stanford University Press.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). *Sample survey methods and theory* (2 vols.). New York: Wiley. (Original work published 1953)
- Harcourt Assessment: <http://www.harcourtassessment.com>
- Harcourt/PsychCorp: http://www.harcourt.com/bu_info/harcourt_assessment.html
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
- Harmonic mean article: http://en.wikipedia.org/wiki/Harmonic_mean
- Harper, H. L. (1974–1976). The method of least squares and some alternatives. Part I, II, IV, V, VI. *International Statistical Review*, 42, 147–174; 42, 235–264; 43, 1–44; 43, 125–190; 43, 269–272; 44, 113–159.
- Harré, R., & Madden, E. H. (1975). *Causal powers*. Oxford, UK: Blackwell.
- Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69, 635–640.
- Harris, D. D. (1963). *Children's drawings as measures of intellectual maturity: A revision and extension of the*

- Goodenough Harris Draw a Man Test. New York: Harcourt, Brace & World.
- Harris, H. (2001). Content analysis of secondary data: A study of courage in managerial decision-making. *Journal of Business Ethics*, 34(3/4), 191–208.
- Harris, M. B., Harris, R. J., & Bochner, S. (1982). Fat, four eyed, and female: Stereotypes of obesity, glasses, and gender. *Journal of Applied Social Psychology*, 12, 503–516.
- Harris, R. J. (1993). Multivariate analysis of variance. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 255–296). New York: Marcel Dekker.
- Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Erlbaum.
- Harter, S. (1985). *Manual for the self-perception profile for children*. Denver, CO: University of Denver.
- Hartigan, J. A., & Kleiner, B. (1984). A mosaic of television ratings. *American Statistician*, 38, 32–35.
- Hartley, J. (1998). *An evaluation of structured abstracts in journals published by the British Psychological Society*. Retrieved from <http://cogprints.org/587/00/199801001.html>
- Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (pp. 16–19). Beverly Hills, CA: Sage.
- Hartz, S. M. (2002). *A Bayesian guide for the Unified Model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.
- Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.
- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.
- Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 5, 121–143.
- Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, 16(3), 354–367.
- Haynes, W. O. (1985). Review of Preschool Language Assessment Instrument, Experimental Edition. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 1190–1192). Lincoln: University of Nebraska Press.
- Hays, W. L. (1973). *Statistics*. New York: Holt, Rinehart and Winston.
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart and Winston.
- Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.
- Hayslip, B., Jr., Francis, J., Radika, L. M., Panek, P. E., & Bosmajian, L. (2002). The development of a scoring system for the Gerontological Apperception Test. *Journal of Clinical Psychology*, 58, 471–478.
- Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, 12, 61–75.
- Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81, 1000–1004.
- Healy, M. J. R. (1968). Multivariate normal plotting. *Applied Statistics*, 17, 157–161.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Heiman, G. W. (2006). *Basic statistics for the behavioral sciences*. Boston: Houghton Mifflin.
- Heise, D. R. (1970). The semantic differential and attitude research. In G. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.
- Helms, J. E. (1983). *Practitioner's guide to the Edwards Personal Preference Schedule*. Springfield, IL: Charles C Thomas.
- Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press.
- Henly, S. J., Klebe, K. J., McBride, J. R., & Cudek, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement*, 13, 363–371.
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Henry A. Murray and Christiana D. Morgan biographies: <http://www.mhhe.com/mayfieldpub/psychtesting/profiles/>
- Henze, N. (2002). Invariant tests for multivariate normality: A critical review. *Statistical Papers*, 43, 467–503.
- Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley.
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.
- Heyde, C. C., & Seneta, C. (Eds.). (2001). *Statisticians of the centuries*. New York: Springer.
- Hieronymus, A. (1973). *Iowa Test of Basic Skills: Manual for administrators, supervisors, and counselors*. Chicago: Houghton Mifflin.

- Hieronymus, A. (1986). *Iowa Test of Basic Skills: Forms G and H*. Chicago: Riverside.
- Higgins, J. J., & Keller-McNulty, S. (1995). *Concepts in probability and stochastic modeling*. Belmont, CA: Duxbury.
- Hilsenroth, M. J., & Segal, D. L. (Eds.). (2004). *Comprehensive handbook of psychological assessment (Volume 2): Personality assessment*. Hoboken, NJ: Wiley.
- Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Orlando, FL: Harcourt Brace Jovanovich.
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician*, 52, 181–184.
- Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, 11, 36–53.
- Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2, 259–278.
- Historiometrics articles: <http://psychology.ucdavis.edu/Simonton/dkspdf.html> (especially publications 204 and 257, which provide overviews)
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–803.
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.
- Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, 27, 324–335.
- Hodges, J. L., Jr., & Lehmann, E. L. (1970). *Basic concepts of probability and statistics* (2nd ed.). San Francisco: Holden-Day.
- Hoehn, L., & Niven, I. (1985). Averages on the move. *Mathematics Magazine*, 58, 151–156.
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19–24.
- Hoffman, A., Field, S., & Sawilowsky, S. (1995). *Self-Determination Knowledge Scale: Forms A & B*. Austin, TX: Pro-Ed.
- Hofmann, H. (2003). Constructing and reading mosaicplots. *Computational Statistics and Data Analysis*, 43, 565–580.
- Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Hoh, J., & Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences USA*, 97(17), 9615–9617.
- Holden, R. R. (1996). *Holden Psychological Screening Inventory (HPSI)*. North Tonawanda, NY: Multi-Health Systems.
- Holden, R. R. (2000). Application of the construct heuristic to the screening of psychopathology: The Holden Psychological Screening Inventory (HPSI). In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 97–121). Boston: Kluwer.
- Holden, R. R., & Jackson, D. N. (1992). Assessing psychopathology using the Basic Personality Inventory: Rationale and applications. In J. Rosen & P. McReynolds (Eds.), *Advances in psychological assessment* (Vol. 8, pp. 165–199). New York: Plenum.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
- Hollands, J. G., & Spence, I. (2001). The discrimination of graphical elements. *Applied Cognitive Psychology*, 15, 413–431.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.
- Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. Chicago: University of Chicago.
- Hopkins, R., Kilik, L., Day, D., Rows, C., & Hamilton, P. (2004). The Revised Kingston Standardized Cognitive Assessment. *International Journal of Geriatric Psychiatry*, 19, 320–326.
- Hopkins, R., Kilik, L., Day, D., Rows, C., & Hamilton, P. (2005). The Brief Kingston Standardized Cognitive Assessment–Revised. *International Journal of Geriatric Psychiatry*, 20, 227–231.
- Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30, 1–15. Retrieved from www.sportsci.org/resource/stats/precision.html
- Horn, J. L., Wanberg, K. W., & Foster, F. M. (1990). *Guide to the Alcohol Use Inventory (AUI)*. Minneapolis, MN: National Computer Systems.
- Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge, UK: Cambridge University Press.
- Hosmer, D., & Lemeshow, S. (2001). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 7, 29–43.
- How to administer a true/false item test using Blackboard: <http://www.mc.maricopa.edu/other/ctl/coursematerial/respondus/importfromword.pdf>
- Howard, P. (1996, February). *Standard error of measurement (SE_M)*. Technical assistance paper, Florida Department of Education, Division of Public Schools, Bureau of Student Services and Exceptional Education. Retrieved from <http://sss.usf.edu/pdf/standerrmeastap.pdf>

- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury.
- Howell, K. K., & Bracken, B. A. (1992). Clinical utility of the Bracken Basic Concept Scale as a preschool intellectual screener: Comparison with the Stanford-Binet for Black children. *Journal of Clinical Child Psychology, 21*, 255–261.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. New York: Chapman & Hall.
- Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 131–137.
- Hsu, L. M. (2000). Effects of directionality of significance tests on the bias of accessible effect sizes. *Psychological Methods, 5*, 333–342.
- Hsu, L. M. (2002). Fail-Safe Ns for one- and two-tailed tests lead to different conclusions about publication bias. *Understanding Statistics, 1*, 85–100.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings for the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–223). Berkeley: University of California Press.
- Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., & Mesbah, M. (Eds.). (2002). *Goodness of fit tests and model validity*. Papers from the International Conference held in Paris, May 29–31, 2000. Boston: Birkhäuser Boston.
- Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 183–208). San Diego, CA: Academic Press.
- Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Allyn & Bacon.
- Huffman, D. A. (1952). A method for the construction of minimum redundancy codes. *Proceedings of the IRE, 40*(9), 1098–1101.
- Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*, 253–294.
- Huitema, B. E. (in preparation). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and observational studies* (2nd ed.). Hoboken, NJ: Wiley.
- Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*, 104–116.
- Hummel, J. (1996). Linked bar charts: Analysing categorical data graphically. *Journal of Computational Statistics, 11*, 23–33.
- Hunter, I. M. L. (1964). *Memory*. London: Penguin.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Hunter, M. A., & May, R. B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Canadian Journal of Experimental Psychology, 57*, 176–188.
- Hurlburt, R. T. (2006). *Comprehending behavioral statistics*. Belmont, CA: Wadsworth.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.
- Hutchinson, J. E. (1981). Fractals and self-similarity. *Indiana University Mathematics Journal, 30*, 713–747.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 86*, 964–973.
- Hwang, F. K. (1977). A generalization of the Karlin-McGregor theorem on coincidence probabilities and an application to clustering. *Annals of Probability, 5*(5), 814–817.
- Hypergeometric distribution articles: <http://mathworld.wolfram.com/HypergeometricDistribution.html> and http://en.wikipedia.org/wiki/Hypergeometric_distribution
- HyperStat Online tutorial and steps for the computation of the variance: <http://davidmlane.com/hyperstat/A29697.html>
- Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*, 497–500.
- IDEA and special education updates on regulations, articles, and other general information: <http://www.ed.gov/offices/OSERS/IDEA/>
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710.
- Immediate and Delayed Memory Tasks description and the laboratory where they were developed: <http://www1.wfubmc.edu/psychiatry/Research/NRLC/>
- Impara, J. C., & Murphy, L. L. (Eds.) (1996). *Buros desk reference: Assessment of substance abuse*. Lincoln, NE: Buros Institute of Mental Measurements.
- Independent and dependent variables further examples and discussion: <http://www.cs.umd.edu/~mstark/exp101/expvars.html>
- Independent variables further definition and examples: http://en.wikipedia.org/wiki/Independent_variable
- Individuals with Disabilities Education Act (IDEA) (20 U.S.C.A. §§ 1400 *et seq.* *L.B. and J.B. v. Nebo School Dist.*, 214 F. Supp.2d 1172 (D. Utah 2002)).
- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105–17, 111 Stat. 37 (codified as amended at 20 U.S.C. § 1400 *et seq.*).

- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 103-218 (GPO 1997).
- Inferential statistics: <http://www.socialresearchmethods.net/kb/statinf.htm>
- Information Referenced Testing research program: <http://www.gseis.ucla.edu/faculty/bruno/rassessment.htm>
- Inselberg, A. (1985). The plane with parallel coordinates. *Visual Computer*, 1, 69–91.
- Inselberg, A. (1999). Don't panic . . . just do it in parallel! *Computational Statistics*, 14, 53–77.
- Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.
- Institute for Personality and Ability Testing information about the test, its underlying theory, and interpretive programs: www.ipat.com
- Instruction on using spreadsheet functions: <http://spreadsheets.about.com/od/excelfunctions/>
- International Comparative Studies in Education: Descriptions of selected large-scale assessments and case studies: http://books.nap.edu/html/icse/study_h.html
- Intervention Central: <http://www.interventioncentral.org>
- Iowa Tests of Basic Skills information: <http://www.education.uiowa.edu/itp/itbs/index.htm>
- Iowa Tests of Educational Development: Guide to research and development*. (2003). Itasca, IL: Riverside.
- Iowa Tests of Educational Development: Interpretive guide for teachers and counselors*. (2001). Itasca, IL: Riverside.
- ISI Web of Knowledge Web site: <http://www.isiwebofknowledge.com/>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109–135.
- J. A. Naglieri Web site: <http://www.mypsychologist.com>
- Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, 96, 589–596.
- Jackson Inventory information: <http://www.sigmaassessmentsystems.com/assessments/jpir.asp>
- Jackson Vocational Interest Survey online: <http://www.jvis.com/>
- Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, 78, 229–248.
- Jackson, D. N. (1994). *Jackson Personality Inventory-Revised manual*. Port Huron, MI: Sigma Assessment Systems.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79–87.
- Jacobs, R. A., Tanner, M. A., & Peng, F. (1996). Bayesian inference for hierarchical mixtures of experts with applications to regression and classification. *Statistical Methods in Medical Research*, 5, 375–390.
- Jacobson, J. O. (2004). Place and attrition from substance abuse treatment. *Journal of Drug Issues*, 34, 23–50.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69(1), 85–98.
- James E. Bruno Research Activities page: <http://www.gseis.ucla.edu/faculty/bruno/brunor.htm>
- Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, 62, 257–270.
- Jankowski, D., & Millon, T. (2002). *A beginner's guide to the MCMI-III*. Washington, DC: American Psychological Association.
- Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14, 491–498.
- Java applet to compute probability values for chi-square: http://psych.rice.edu/online_stat/analysis_lab/chi_square_prob.html
- Java applet simulation of the chi-square test of association: http://www.ruf.rice.edu/~lane/stat_sim/contingency/index.html
- Java applets ("Piface"): www.stat.uiowa.edu/~rlenth/Power/ (Either these may be run in a browser with appropriate Java plug-in, or they may be downloaded to run offline. This page also provides links to a number of other online power calculators.)
- Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, 10, 93–98.
- Jensen, A.-M., Knutsen, T., & Skonhøft, A. (Eds.). (2003). *Visiting Malthus: The man, his times, the issues*. Copenhagen: Copenhagen Business School Press.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jensen, A. R. (2003). Regularities in Spearman's law of diminishing returns. *Intelligence*, 31, 95–105.
- Jensen, A. R., & Rohwer, W. D., Jr. (1966). The Stroop color-word test: A review. *Acta Psychologica*, 25, 36–93.
- Jewell, N. P. (2004). *Statistics for epidemiology*. New York: Chapman & Hall/CRC.
- Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology*, 150(4), 327–333.
- Johansson, C. B. (1986). *Manual for the Career Assessment Inventory: The enhanced version*. Minneapolis, MN: National Computer Systems.
- John Leonard Horn: <http://www.usc.edu/projects/nexus/faculty/dept-ldsg/hornjohn/horn.shtml>
- Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriousness and spuriousness: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153–162.

- Johnson, D. E. (1998). *Applied multivariate methods for data analysts*. Pacific Grove, CA: Duxbury.
- Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Johnson, S. M. (1996). *The practice of emotionally focused marital therapy: Creating connections*. New York: Brunner/Mazel.
- Johnson, S. M., Hunsley, J., Greenberg, L., & Schindler, D. (1999). Emotionally focused couples therapy: Status and challenges. *Clinical Psychology: Science & Practice*, 6, 67–79.
- Jonathan Barzilai Web page: <http://myweb.dal.ca/barzilai/>
- Jones, J. H. (1993). *Bad blood: The Tuskegee syphilis experiment*. New York: Free Press. (Originally published in 1981)
- Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181–214.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Seminar Press.
- Journal of the American Statistical Association* and its predecessors are archived on the JSTOR Web site: www.jstor.org
- Journal of the American Statistical Association* and other ASA publications information: www.amstat.org
- Journal of Applied Behavior Analysis*: <http://seab.envmed.rochester.edu/jaba/index.html>
- Journal of Official Statistics* online: <http://www.jos.nu> (Click on “Search,” and then search keyword “weight.” A number of articles will be found on topics such as noncoverage, raking, and other advanced topics.)
- Journal of Statistics Education* home page: <http://www.amstat.org/publications/jse/>
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.
- Juni, S. (1979). Theoretical foundations of projection as a defence mechanism. *International Review of Psychoanalysis*, 6, 115–130.
- Juni, S., & Koenig, E. J. (1982). Contingency validity as a requirement in forced-choice item construction: A critique of the Jackson Vocational Interest Survey. *Measurement and Evaluation in Guidance*, 14(4), 202–207.
- Junker, B., & Sijsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Kagan, A. M., Linnik, Y. V., & Rao, C. R. (1973). *Characterization problems in mathematical statistics*. New York: Wiley.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kalbfleisch, J. D., & Prentice, R. L. (2002). *Statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.
- Kamler, B., & Thomson, P. (2004). Driven to abstraction: Doctoral supervision and writing pedagogies. *Teaching in Higher Education*, 9(2), 195–209.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Kane, M. T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.
- Kaplan, R. M., & Anderson, J. P. (1996). The general health policy model: An integrated approach. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials* (2nd ed., pp. 309–322). Philadelphia: Lippincott-Raven.
- Kaplan, R. M., Sieber, W. J., & Ganiats, T. G. (1997). The Quality of Well-Being Scale: Comparison of the interviewer-administered version with a self-administered questionnaire. *Psych Health*, 12, 783–791.
- Karelitz, S., Fisichelli, V. R., Costa, J., Karelitz, R., & Rosenfeld, L. (1964). Relation of crying activity in early infancy to speech and intellectual development at age three years. *Child Development*, 35, 769–777.
- Karson, M., Karson, S., & O'Dell, J. (1997). *16PF interpretation in clinical practice: A guide to the fifth edition*. Champaign, IL: Institute for Personality and Ability Testing.
- Karson, M., & Kline, C. (2004, April 4). Two interpretations of Jim Wood's specimen Rorschach protocol. *WebPsychEmpiricist*. Retrieved from <http://home.earthlink.net/~rkmck/vault/karson04/karson04.pdf>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90, 773–795.
- Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.
- Katz, J. A. (1992). Secondary analysis in entrepreneurship: An introduction to databases and data management. *Journal of Small Business Management*, 30(2), 74–75.

- Katz, S., Branch, L. G., Branson, M. H., Papsidero, J. A., Beck, J. C., & Greer, D. S. (1983). Active life expectancy. *New England Journal of Medicine*, 309, 1218–1224.
- Kaufman, A. S. (1978). The importance of basic concepts in the individual assessment of preschool children. *Journal of School Psychology*, 16, 208–211.
- Kaufman, A. S., & Lichtenberger, E. O. (2005). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.
- Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of K-ABC-II assessment*. Circle Pines, MN: American Guidance Service.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: Wiley.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- Kazdin, A. E. (1998). The case study and single-case research designs. In A. E. Kazdin (Ed.), *Research design in clinical psychology* (3rd ed., pp. 202–244). Needham Heights, MA: Allyn & Bacon.
- Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20, 643–656.
- Kelly, H., Riddell, M. A., Gidding, H. F., Nolan, T., & Gilbert, G. L. (2002). A random cluster survey and a convenience sample give comparable estimates of immunity to vaccine preventable diseases in children of school age in Victoria, Australia. *Vaccine*, 20(25–26), 3130–3136.
- Kendall, M. G. (1955). *Rank correlation methods*. New York: Hafner.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.
- Kendall, M. G., & Stuart, A. (1966). *The advanced theory of statistics*, Vol. 3. New York: Hafner.
- Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy & Research*, 11, 289–299.
- Kenney, J. F., & Keeping, E. S. (1962). Harmonic mean. *Mathematics of statistics*, Pt. 1 (3rd ed., pp. 57–58). Princeton, NJ: Van Nostrand.
- Kenney, J. F., & Keeping, E. S. (1962). *Mathematics of statistics: Part I* (3rd ed.). Princeton, NJ: Van Nostrand.
- Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine*, 17, 2723–2732.
- Keppel, G., Saufley, W., & Tokunaga, H. (1992). *Introduction to design and analysis: A student's handbook*. New York: W. H. Freeman.
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
- Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.
- Kerns, R. D., & Rosenberg, R. (1995). Pain-relevant responses from significant others: Development of a significant-other version of the WHYMPI scales. *Pain*, 61, 245–249.
- Kerns, R. D., Turk, D. C., & Rudy, T. E. (1985). The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). *Pain*, 23, 345–356.
- Kherif, F., Poline, J.-P., Mériaux, S., Benali, H., Flandin, G., & Brett, M. (2003). Group analysis in functional neuroimaging: Selecting subjects using similarity measures. *NeuroImage*, 20, 2197–2208.
- Kim, P. J., & Jennrich, R. I. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion, D_{mn} , $m \leq n$. In H. L. Harter & D. B. Owen (Eds.), *Selected tables in mathematical statistics*, Vol. 1. Providence, RI: American Mathematical Society.
- Kingsbury, N., Droitcour, J. A., & Larson, E. M. (2003). *Estimating illegal immigrants in a census or survey: The three-card method—an innovative technique*. Conference of European Statisticians, Joint ECE-Eurostat Work Session on Migration Statistics, U.N. Statistics Division, Working paper No. 3. Retrieved from <http://www.unecce.org/stats/documents/2003/04/migration/wp.3.e.pdf>
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational & Psychological Measurement*, 61, 213–218.
- Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Kirschenbaum, M. G. Electronic theses and dissertations in the humanities: A directory of online references and resources: <http://etext.virginia.edu/ETD/>
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Klebanov, L. B. (1976). A general definition of unbiasedness. *Theory of Probability and Its Application*, 21, 571–585.
- Kleinbaum, D., Kupper, L., Muller, K., & Nizam, A. (1998). *Applied regression analysis and multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury.
- Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Kline, T. J. B., & Dunn, B. (2000). Analysis of interaction terms in structural equation models: A non-technical demonstration using the deviation score approach. *Canadian Journal of Behavioural Science*, 32, 127–132.

- Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin*, 85, 410–416.
- Knaub, J. R., Jr. (1997). *Weighting in regression for use in survey methodology*. Retrieved April 11, 2006, from <http://interstat.statjournals.net/>
- Knoff, H. M., & Prout, H. T. (1985). *Kinetic drawing system for family and school: A handbook*. Los Angeles: Western Psychological Services.
- Knowledge Factor Web site: <http://www.knowledgefactor.com/>
- Knuth, D. E. (1997). *The art of computer programming, Volume 2: Seminumerical algorithms* (3rd ed.). Reading, MA: Addison-Wesley.
- Kolar, J. C., & Salter, E. M. (1997). Craniofacial anthropometry. Practical measurement of the head and face for clinical, surgical and research use. Springfield, IL: Charles C Thomas.
- Kolen, M. J. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices*. New York: Springer.
- Koocher, G. P., & Rey-Casserly, C. M. (2002). Ethical issues in psychological assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of assessment psychology*. New York: Wiley.
- Korpiva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Korth, B. A., & Tucker, L. R. (1976). The distribution of chance congruence coefficients from simulated data. *Psychometrika*, 40, 361–372.
- Korth, B., & Tucker, L. R. (1976). Procrustes matching by congruence coefficients. *Psychometrika*, 41, 531–535.
- Kovaleski, J., & Prasse, D. P. (2004). Response to instruction in the identification of learning disabilities: A guide for school teams. *NASP Communiqué*. Retrieved from <http://www.nrcl.org/html/research/rti/RTIinfo.pdf>
- Kraemer, H. C., & Thieman, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
- Kroner, D. G., Black, E. R., & Reddon, J. R. (1996). Item content of the Basic Personality Inventory. *Multivariate Experimental Clinical Research*, 11, 61–73.
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Beimer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and quality* (pp. 141–164). Thousand Oaks, CA: Sage.
- Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, 56, 16–26.
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27, 313–376.
- Kuder, F. (1977). *Activity interests and occupational choice*. Chicago: Science Research Associates.
- Kuder, F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey general manual* (3rd ed.). Monterey, CA: CTB/McGraw-Hill.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Kuder Career Planning System: <http://www.kuder.com/>
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 188–229.
- Kuhn, T. (1962). *The structure of scientific revolution*. Chicago: University of Chicago Press.
- Kuhn, T. (1974). *The structure of scientific revolution* (2nd ed.). Chicago: University of Chicago Press.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127, 162–181.
- Kurt Thearling's data mining page: <http://www.thearling.com/>
- Kury, S. P., Rodrigue, J. R., & Peri, M. G. (1998). Smokeless tobacco and cigarettes: Differential attitudes and behavioral intentions of young adolescents toward a hypothetical new peer. *Journal of Clinical Child Psychology*, 27(4), 415–422.
- Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). Chicago: McGraw-Hill Irwin.
- Kvalseth, T. O. (1989). Note on Cohen's kappa. *Psychological Reports*, 65, 223–226.
- Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Norwell, MA: Kluwer.
- Laboratory Behavioral Measures of Impulsivity tasks and the laboratory where they were developed: <http://www1.wfubmc.edu/psychiatry/Research/NRLC/>
- Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222–230.
- Lane, D. (2003). *Levels of measurement*. Retrieved from <http://cnx.rice.edu/content/m10809/latest/>
- Lane, D. M., Peres, S. C., Sándor, A., & Napier, H. A. (2005). A process for anticipating and executing icon selection in graphical user interfaces. *International Journal of Human Computer Interaction*, 19, 241–252.
- Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5, 1–18.
- Lange, K. (1999). *Numerical analysis for statisticians*. New York: Springer.
- LaPointe, A. E., Mead, N. A., & Phillips, G. W. (1989). *A world of differences: An international assessment of*

- mathematics and science*. New Jersey: Educational Testing Service.
- Larsen, M. D., & Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32–41.
- Laughlin, T. (1995). The school readiness composite of the Bracken Basic Concept Scale as an intellectual screening instrument. *Journal of Psychoeducational Assessment*, 13, 294–302.
- Lavit, C., Escoufier Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS) method. *Computational Statistics & Data Analysis*, 18, 97–119.
- Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed.). New York: McGraw-Hill.
- Law School Admission Council: <http://www.lsac.org/>
- Lee, M. (1996). *Methods of moments and semiparametric econometrics for limited dependent variable models*. New York: Springer.
- Lee, M. L. T. (1966). Properties and applications of the Sarmonov family of bivariate distributions. *Communications in Statistics*, 25, 1207–1222.
- Lee, R. M., & Renzetti, C. M. (1993). The problem of researching sensitive topics: An overview and introduction. In C. M. Renzetti & R. M. Lee (Eds.), *Researching sensitive topics*. Newbury Park, CA: Sage.
- Lee Cronbach's two most influential papers: <http://psychclassics.yorku.ca/Cronbach/construct.htm> and <http://psychclassics.yorku.ca/Cronbach/Disciplines>
- Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
- Leemis, L. M., & Park, S. K. (2006). *Discrete-event simulation: A first course*. Englewood Cliffs, NJ: Prentice Hall.
- Leff, G. (1975). *William of Ockham: The metamorphosis of Scholastic discourse*. Manchester, UK: Manchester University Press.
- Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd English ed.). Amsterdam: Elsevier Science.
- Lehmann, E. L. (1951). A general concept of unbiasedness. *Annals of Mathematical Statistics*, 22, 587–592.
- Lehmann, E. L. (1983). Estimation with inadequate information. *Journal of the American Statistical Association*, 78, 624–627.
- Lehmann, E. L. (1983). *Theory of point estimation*. New York: Wiley.
- Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.
- Lehmann, E. L. (1991). *Theory of point estimation*. Pacific Grove, CA: Wadsworth & Brooks/Cole.
- Lehmann, E. L. (2004). Optimality and symposia: Some history. In J. Rojo & V. Perez-Abreu (Eds.), *The First Erich L. Lehmann Symposium: Optimality*. IMS Lecture Notes and Monograph Series, Vol. 44.
- Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer-Verlag.
- Lehmann, I. J. (1990). The state of NCME: Remembering the past, looking to the future. *Educational Measurement: Issues and Practice*, Spring, 3–10.
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician*, 55, 187–193.
- Leon, S. J. (1998). *Linear algebra with applications* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Leppin, V. (2003). *Wilhelm von Ockham: Gelehrter, Streiter, Bettelmönch*. Darmstadt, Germany: Wissenschaftliche Buchgesellschaft.
- Leung, M.-Y., Choi, K. P., Xia, A., & Chen, L. H. Y. (2005). Nonrandom clusters of palindromes in herpesvirus genomes. *Journal of Computational Biology*, 12(3), 331–354.
- Leung, M.-Y., Schachtel, G. A., & Yu, H.-S. (1994). Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. *Nonlinear World*, 1, 445–471.
- Levant, R. F., & Fischer, J. (1998). The Male Role Norms Inventory. In C. M. Davis, W. H. Yarber, R. Bauserman, G. Schreer, & S. L. Davis (Eds.), *Sexuality-related measures: A compendium* (2nd ed., pp. 469–472). Thousand Oaks, CA: Sage.
- Levant, R. F., & Richmond, K. (n.d.). *A program of research on masculinity ideologies using the Male Role Norms Inventory*. Manuscript submitted for publication.
- Levels of Measurement page from Research Methods Knowledge Base by W. M. Trochim: <http://www.socialresearchmethods.net/kb/measlevl.htm>
- Levene, H. (1960). Robust test for the equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press.
- Lévy Véhel, J., & Lutton, E. (Eds.). (2005). *Fractals in engineering: New trends in theory and applications*. London: Springer.
- Lévy Véhel, J., Lutton, E., & Tricot, C. (Eds.). (1997). *Fractals in engineering*. London: Springer.
- L'Hermier des Plantes, H. (1976). *Structuration des tableaux à trois indices de la statistique*. Thèse de troisième cycle, Université de Montpellier, France.
- Li, C. C. (1975). *Path analysis: A primer*. Pacific Grove, CA: Boxwood Press.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Society*, 86, 316–342.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Lichtenberger, E. O., & Smith, D. B. (2005). *Essentials of WIAT-II and KTEA-II assessment*. Hoboken, NJ: Wiley.
- Life values inventory: <http://www.lifevaluesinventory.com>
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Likert scales and detecting response sets (e.g., acquiescence bias): <http://www.rasch.org/erp9.htm>
- Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika*, 11, 46–54.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA.
- Linn, R. L. (1995). High stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In T. Oaklund & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 49–74). Boston: Kluwer.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, 13(1), 5–8, 15.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Upper Saddle River, NJ: Prentice Hall.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Applied Social Research Methods Series Vol. 49). Thousand Oaks, CA: Sage.
- LISREL student version (8.7) for downloading: <http://www.ssicentral.com/index.html> (Note that there are limitations on the size of the model and no technical support.)
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. (1996). *SAS system for mixed models*. SAS Institute.
- Little, B. R. (1983). Personal projects analysis: A rationale and method for investigation. *Environment and Behavior*, 15, 273–309.
- Little, B. R. (1989). Trivial pursuits, magnificent obsessions and the search for coherence. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions*. New York: Springer-Verlag.
- Little, B. R. (2005). Personality science and personal projects: Six impossible things before breakfast. *Journal of Research in Personality*, 39, 4–21.
- Little, B. R., Salmela-Aro, K., & Phillips, S. D. (Eds.). (in press). *Personal project pursuit: Goals, action and human flourishing*. Mahwah, NJ: Erlbaum.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Liu, A. (2002). Efficient estimation of two seemingly unrelated regression equations. *Journal of Multivariate Analysis*, 82, 445–456.
- Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633–648.
- Liu, C., Rubin, D. B., & Wu, Y. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* 85, 755–770.
- Liu, Z. Y., & Xu, L. (2003). Topological local principal component analysis. *Neurocomputing*, 55, 739–745.
- Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage.
- Loglinear analysis is part of the advanced modules added to the basic procedures of SPSS: http://www.spss.com/advanced_models/data_analysis.htm
- Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher*, 22, 12–23.
- Lohman, D. F. (2005). The role of non-verbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly*, 49, 111–138.
- Lohman, D. F. (in press). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*.
- Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.
- Lombroso, C. (1876). *l'uomo delinquente*. Milan, Italy: Hoepli.
- Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *American Statistician*, 49, 64–70.
- Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 20, 1–22.
- Lord, F. M. (1963). Biserial estimates of correlation. *Psychometrika*, 28, 81–85.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper & Row.
- Lord, M. F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64.
- Lorge, I., & Thorndike, R. L. (1954). *The Lorge-Thorndike Intelligence Tests*. Boston: Houghton Mifflin.
- Louw, A. E., & Ramkissoon, S. (2002). The suitability of the Roberts Apperception Test for Children (RATC), The House-Tree-Person (H-T-P), and Draw-a-Person (D-A-P) scales in the identification of child sexual abuse in the Indian community: An exploratory study. *Southern African Journal of Child and Adolescent Mental Health*, 14, 91–106.
- Lubin, B., Zuckerman, M., Hanson, P. G., Armstrong, T., Rinck, C. M., & Seever, M. (1986). Reliability and validity of the Multiple Affect Adjective Check List-Revised. *Journal of Psychopathology and Behavioral Assessment*, 8(2), 103–117.

- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology? *Journal of Extension*, 35(5). Retrieved December 13, 2001, from <http://www.joe.org/joe/1997october/tt2.html>
- Luo, S., & Klohnen, E. C. (2005). Assortative mating and marital quality in newlyweds: A couple-centered approach. *Journal of Personality & Social Psychology*, 88, 304–326.
- Lutkepohl, H. (2005). *New introduction to multiple time series*. New York: Springer.
- Lynn, P., & Jowell, R. (1996). How might opinion polls be improved? The case for probability sampling. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 159, 21–28.
- Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.
- Maassen, G. H., & Bakker, A. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods & Research*, 30, 241–270.
- MacDonald, S. (1997). *The portfolio and its use: A road map for assessment*. Beltsville, MD: Gryphon House.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley and Los Angeles: University of California Press.
- MacTutor History of Mathematics archive: <http://turnbull.mcs.st-and.ac.uk/~history/>
- Mager, R. F. (1984). *Preparing instructional objectives* (2nd ed.). Belmont, CA: David S. Lake.
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.
- Mallat, S. (1999). *A wavelet tour of signal processing*. San Diego, CA: Academic Press.
- Mandelbrot, B. (1983). *The fractal geometry of nature*. San Francisco: W. H. Freeman.
- Mandelbrot, B., & van Ness, J. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10, 422–437.
- Manet and Mondrian data analysis tools with extensive functionality for mosaic plots in an interactive framework: <http://www.rosuda.de/software>
- Manetti, M., Schneider, B. H., & Siperstein, G. N. (2001). Social acceptance of children with mental retardation: Testing the contact hypothesis with an Italian sample. *International Journal of Behavioral Development*, 25, 279–286.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences* (167–171). Monterey, CA: Brooks/Cole.
- Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: Freeman.
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.
- Maris, E. (1995). Psychometric latent response models. *Psychometrika*, 60, 523–547.
- Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73(361), 194–196.
- Markov, A. A. (1948). Izbrannye trudy po teorii nepre-ryvnyh drobei i teorii funktsii naimeney uklonyayushchih-sya ot nulya. N. I. Ahiezer (Biograficheskii ocherk I primechaniya), & L. I. Volkobyskii. Moskva, Leningrad, SSSR: OGIZ, Gosudarstvennoe izdatel'stvo tekhniko-teoreticheskoi literatury. [Selected works on theory of continuous fractions and theory of the least deviated from zero functions (N. I. Ahiezer, Ed., & L. I. Volkobyskii, Trans., selected chapters). Moscow, Leningrad, USSR: OGIZ, State Publishing House for Technical-Theoretical Literature.]
- Marks, L. E., & Gescheider, G. A. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 91–138). New York: Wiley.
- Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research*, 45, 7–34.
- Marsaglia, G. (2003). Random number generators. *Journal of Modern Applied Statistical Methods*, 2(1), 2–13.
- Mart, E. G. (1999). Problems with the diagnosis of factitious disorder by proxy in forensic settings. *American Journal of Forensic Psychology*, 17(1), 69–82.
- Martens, H., & Naes, T. (1989). *Multivariate calibration*. London: Wiley.
- Martin, C. G., & Games, P. A. (1977). ANOVA tests for homogeneity of variance: Non-normality and unequal samples. *Journal of Educational Statistics*, 2, 187–206.
- Martinez, W. L., & Martinez, A. R. (2001). *Computational statistics handbook with Matlab*. New York: Chapman & Hall.
- Mason, M. (1999). A review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement and Evaluation in Counseling and Development*, 32, 111–118.

- Matheny, K. B., Aycok, D., Curlette, W. L., & Junker, G. (1993 November). Coping Resources Inventory for Stress: A measure of perceived coping resources. *Journal of Clinical Psychology*, 49(6), 815–830.
- Matheny, K. B., & Curlette, W. L. (1998). The Coping Resources Inventory for Stress: A comprehensive measure of stress-coping resources. In C. P. Zarakett & R. J. Wood (Eds.), *Evaluating stress: A book of resources*. Lanham, MD: Scarecrow.
- Matheny, K. B., Curlette, W. L., Aysan, F., Herrington, A., Gfroerer, C. A., Thompson, D., et al. (2002). Coping resources, perceived stress, and life satisfaction among Turkish and American university students. *International Journal of Stress Management*, 9(2), 81–97.
- Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ-III tests of achievement assessment*. New York: Wiley.
- Maxwell, S. E., and Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- May, W. L., & Johnson, W. D. (1997). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine*, 16, 2127–2136.
- May, W. L., & Johnson, W. D. (2001). Symmetry in square contingency tables: Tests of hypotheses and confidence interval construction. *Journal of Biopharmaceutical Statistics*, 11(1–2), 23–33.
- McAllister, L. W. (1996). *A practical guide to CPI interpretation* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
- McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, 38(1), 115–142.
- McArthur, D. S., & Roberts, G. E. (1982). *Roberts Apperception Test for Children (RATC) manual*. Los Angeles: Western Psychological Services.
- McCallum, R. S., & Bracken, B. A. (2005). The Universal Nonverbal Intelligence Test: A multidimensional measure of intelligence. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 425–440). New York: Guilford.
- McCallum, R. S., Bracken, B. A., & Wasserman, J. (2001). *Essentials of nonverbal assessment*. New York: Wiley.
- McCaulley, M. H. (2000). Myers-Briggs Type Indicator: A bridge between counseling and consulting. *Consulting Psychology Journal: Practice and Research*, 52, 117–132.
- McCord Adams, M. (1970). Intuitive cognition, certainty, and scepticism in William Ockham. *Traditio*, 26, 389–398.
- McCord Adams, M. (1987). *William Ockham* (Vols. 1–2). Notre Dame, IN: University of Notre Dame Press.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
- McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.
- McDonald, R. P. (1970). The theoretical foundations of common factor analysis, principal factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1–21.
- McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *American Statistician*, 32, 12–16.
- McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage*, 23, 250–263.
- McKean, J. W., & Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *American Statistician*, 48, 220–229.
- McKnight, S., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze an intervention time series model with autoregressive error terms. *Psychological Methods*, 5, 87–101.
- McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.
- McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions* (Wiley Series in Probability and Statistics). New York: Wiley.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- McLaughlin, M. E. (1999). Controlling method effects in self-report instruments. *Research Methods Forum*, 4. Retrieved May 17, 2006, from http://www.aom.pace.edu/rmd/1999_RMD_Forum_Method_Effects_in_Self-Reports.htm
- MCMC preprints page: <http://www.statslab.cam.ac.uk/~mcmc/>
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157.
- McNemar table/Excel spreadsheet: <http://blake.montclair.edu/~koppeln/mcnemar.htm> (Using Microsoft Excel's BINOMDIST function, exact p values of the McNemar test are obtained for all $n \leq 1,000$. A table of critical values for McNemar's test statistic M , along with their associated exact p values, was developed for all $n \leq 1,000$ using various traditional significance levels α for both one- and two-tailed tests. The corresponding spreadsheet can be downloaded from this site.)

- McNicol, D. (1972). *A primer of signal detection theory*. London: Allen & Unwin.
- McQuarrie, A. D. R., & Tsai, C.-L. (1998). *Regression and time series model selection*. River Edge, NJ: World Scientific.
- McSweeney, A. J. (1978). Effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis*, 11(1), 47–51.
- Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, 77(4), 531–552.
- Means and medians graphical comparison applets: http://www.ruf.rice.edu/~lane/stat_sim/descriptive/ and <http://standards.nctm.org/document/eexamples/chap6/6.6/>
- Mecklin, C. J., & Donnelly, R. G. (2005). Powerball, expected value, and the law of (very) large numbers. *Journal of Statistics Education*, 13(2). Available at: <http://www.amstat.org/publications/jse/v13n2/mecklin.html>
- Mecklin, C. J., & Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72, 123–138.
- Mecklin, C. J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75, 93–108.
- Medrich, E. A., & Griffith, J. E. (1992, January). *International mathematics and science assessment: What have we learned?* (National Center for Education Statistics, Research and Development Report). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement. Retrieved from <http://nces.ed.gov/pubs92/92011.pdf>
- Meilijson, I. (1989). A fast improvement of the EM algorithm in its own terms. *Journal of the Royal Statistical Society, Series B*, 51, 127–138.
- Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly*, 14(3), 31–36.
- Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2006). *Introduction to probability and statistics* (12th ed.). Pacific Grove, CA: Brooks/Cole.
- Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the two-way repeated measure design. *Multivariate Behavioral Research*, 9, 165–178.
- Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267–278.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Method of maximum likelihood: http://en.wikipedia.org/wiki/Maximum_likelihood
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335–341.
- Meyer, P., & Davis, S. (1992). *The CPI applications guide*. Palo Alto, CA: Consulting Psychologists Press.
- Meyners, M., Kunert, J., & Qannari, E. M. (2000). Comparing generalized procrustes analysis and STATIS. *Food Quality and Preference*, 11, 77–83.
- Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13, 307–336.
- Michailidis, G., & de Leeuw, J. (2000). Multilevel homogeneity analysis with differential weighting. *Computational Statistics and Data Analysis*, 32, 411–442.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Mahwah, NJ: Erlbaum.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–384.
- Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, 10(5), 639–667.
- Michigan State University Extension. (1994, October). *Delphi technique*. Issue Identification Information – III00006. Retrieved November 28, 2001, from <http://www.msue.msu.edu/msue/imp/modii/iii00006.htm>
- Miles, A. S., Russo, C. J., & Gordon, W. M. (1991). The reasonable accommodations provisions of the Americans with Disabilities Act. *Education Law Reporter*, 69(1), 1–8.
- Millar, G. W. (1995). *The creativity man: An authorized biography*. Norwood, NJ: Ablex.
- Miller, C. (2003). Ethical guidelines in research. In J. C. Thomas & M. Herson (Eds.), *Understanding research in clinical and counseling psychology* (pp. 271–293). Mahwah, NJ: Erlbaum.
- Miller, M. D., & Linn, R. L. (2000). Validation of performance assessments. *Applied Psychological Measurement*, 24(4), 367–378.
- Miller, R. B., & Wright, D. (1995). Correction for attrition bias in longitudinal analyses. *Journal of Marriage and the Family*, 57, 921–929.
- Miller, R. G. (1968). Jackknifing variances. *Annals of Mathematical Statistics*, 39, 567–582.
- Miller, R. G. (1986). *Beyond ANOVA: Basics of applied statistics*. New York: Wiley.

- Millon, T. (1990). *Toward a new personology*. New York: Wiley.
- Millon, T., Green, C. J., & Meagher, R. B. (1979). The MBHI: A new inventory for the psychodiagnostician in medical settings. *Professional Psychology: Research and Practice*, 10, 529–539.
- Millon Clinical Multiaxial Inventory-III: <http://home-town.net/MCMI.htm>
- Minnesota Multiphasic Personality Inventory article: <http://en.wikipedia.org/wiki/MMPI>
- Mirkin, B. (2005). Clustering for data mining: A data recovery approach. London: Chapman & Hall.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Mitra, A. (2005). *Fundamentals of quality control and improvement* (2nd ed.). Mason, OH: Brooks/Cole.
- MM ex rel. DM v. School Dist. of Greenville County*, 303 F.3d 523 [169 Educ. L. Rep. 59] (4th Cir. 2002).
- MMPI-2 information site: <http://www.mmpi-info.com/mmpistart.html>
- Model selection criteria PDF presentations by the author, with an emphasis on AIC and its corrected variants: http://www.myweb.uiowa.edu/cavaaugh/ms_seminar.html
- Moivre, A. de (1718). *The doctrine of chance: Or, a method of calculating the probability of events in play*. London: Pearson.
- Molden, D. C., & Dweck, C.-S. (2006). Finding “meaning” in psychology: A lay theories approach to self-regulation, social perception, and social development. *American Psychologist*, 61(3), 192–203.
- Molenberghs, G., & Verbeke, G. (2005). *Discrete longitudinal data*. New York: Springer.
- Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Mastitis in dairy cattle: Local influence to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, 37, 93–113.
- Molin, P., & Abdi, H. (1998). *New tables and numerical approximation for the Kolmogorov-Smirnov/ Lilliefors/ Van Soest test of normality*. Technical report, University of Bourgogne. Available from www.utd.edu/~herve/MA_Lilliefors98.pdf
- Mood, A. M. (1950). *Introduction to the theory of statistics*. New York: McGraw-Hill.
- Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.
- Mooney, D., & Swift, R. (1999). *A course in mathematical modeling*. Washington, DC: MAA.
- Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. New York; W. H. Freeman. (Original source: Occupational Mortality: The Registrar General’s Decennial Supplement for England and Wales, 1970–1972, Her Majesty’s Stationery Office, London, 1978)
- Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: W. H. Freeman.
- Moos, R. (2001). *The Family Environment Scale: An annotated bibliography* (3rd ed.). Redwood City, CA: Mind Garden.
- Moos, R. (2003). *The Social Climate Scales: A user’s guide* (3rd ed.). Redwood City, CA: W. H. Mind Garden.
- Moos, R., & Lemke, S. (1996). *Evaluating residential facilities*. Thousand Oaks, CA: Sage.
- Moos, R., & Moos, B. (1994). *Family Environment Scale manual* (3rd ed.). Redwood City, CA: Mind Garden.
- Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychology Assessment Resources.
- Morey, L. C. (1996). *An interpretive guide to the Personality Assessment Inventory*. Odessa, FL: Psychology Assessment Resources.
- Morey, L. C. (2003). *Essentials of PAI assessment*. Hoboken, NJ: Wiley.
- Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). *SPSS for basic statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
- Morrison, D. F. (1967). *Multivariate statistical methods*. New York: McGraw-Hill.
- Mosaic plots and SAS macros: <http://www.math.yorku.ca/SCS/friendly.html#mosaics>
- Moses, J. A., Jr., Golden, C. J., Ariel, R., & Gustavson, J. L. (1983). *Interpretation of the Luria-Nebraska Neuropsychological Battery, Volume 1*. New York: Grune & Stratton.
- Mosher, D. L. (1968). Measurement of guilt by self-report inventories. *Journal of Consulting and Clinical Psychology*, 32, 690–695.
- Moskowitz, H. R. (1983). *Product testing and sensory evaluation of foods: Marketing and R&D approaches*. Westport, CT: Food & Nutrition Press.
- Mostert, M. P. (2001). Characteristics of meta-analyses reported in mental retardation, learning disabilities, and emotional and behavioral disorders. *Exceptionality*, 9(4), 199–225.
- Moving averages: <http://www.investopedia.com/university/movingaverage/default.asp>
- Mroz, T. A. (1987). The sensitivity of an empirical model of married women’s hours of work to economic and statistical assumptions. *Econometrica*, 55(4), 765–799.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.
- Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.
- Multi-collinearity—Variance inflation and orthogonalization in regression: <http://creative-wisdom.com/computer/sas/collinear.html>
- Multidimensional Aptitude Battery-II: <http://www.sigmaassessmentsystems.com/assessments/mab.asp>

- Multiple Comparison Procedures outline, by Gerald Dallal: <http://www.tufts.edu/~gdallal/mc.htm>
- Multiple imputation software: <http://www.stat.psu.edu/~jls/misoftwa.html>
- Multitrait-Multimethod Matrix: <http://www.socialresearchmethods.net/kb/mtmmmat.htm>
- Murphy, K. R. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Erlbaum.
- Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (2002). *Tests in print VI*. Lincoln: University of Nebraska Press.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Murray, H. A. (1943). *Thematic Apperception Test: Manual*. Cambridge, MA: Harvard University Press.
- Myers & Briggs Foundation: <http://www.myersbriggs.org>
- Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.
- Myers, R. A. (2000). *Classical and modern regression with applications*. Boston: Duxbury.
- Naglieri, J. A. (1985). *Matrix Analogies Test—Expanded Form*. San Antonio, TX: The Psychological Corporation.
- Naglieri, J. A. (1985). *Matrix Analogies Test—Short Form*. San Antonio, TX: The Psychological Corporation.
- Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. San Antonio, TX: Psychological Corporation.
- Naglieri, J. A. (2003). Naglieri Nonverbal Ability Tests: NMAT and MAT-EF. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 175–189). New York: Kluwer Academic/Plenum.
- Nancy Bayley biography: <http://www.webster.edu/~woolfm/bayley.html>
- National Center for Biotechnology Information Human Genome Resources: <http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/>
- National Center for Biotechnology Information Social Analysis of Gene Expression Tag to Gene Mapping: www.ncbi.nlm.nih.gov/SAGE/
- National Center for Biotechnology Information Statistics of Sequence Similarity Scores: www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics standards for mathematics: <http://standards.nctm.org/index.htm>
- National Council on Measurement in Education Web site: ncme.org
- The National Institute for Standards and Technology's guidelines for true and pseudorandom number generation and testing: <http://csrc.nist.gov/rng/>
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.
- National Science Foundation Web site: www.nsf.gov
- National standards for health care: <http://www.hhs.gov/ocr/hipaa/>
- Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60, 532–538.
- Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society—Series A*, 135(3), 370–384.
- Nelson, M., & Gailly, J.-L. (1996). *The data compression book* (2nd ed.). New York: M&T Books.
- Neonatal Behavioral Assessment Scale and T. Berry Brazelton description: <http://www.brazelton-institute.com>
- Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford, UK: Oxford University Press.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, pp. 954–959.
- Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). UCI Repository of machine learning databases. Irvine: University of California, Department of Information and Computer Science. Retrieved from <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*, 231, 289–337.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- Nightingale, F. (1858). *Notes on matters affecting the health, efficiency, and hospital administration of the British Army, founded chiefly on the experience of the late war*. London: Harrison and Sons.
- Niolon, R. (2003). *Notes on projective drawings*. Retrieved August 20, 2005, from http://www.psychpage.com/projective/proj_draw_notes.html
- Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: Toronto University Press.
- NIST/SEMATECH e-Handbook of Statistical Methods: <http://www.itl.nist.gov/div898/handbook/>
- NIST/SEMATECH. (n.d.). Quantile-quantile plot. *NIST/SEMATECH e-Handbook of statistical methods*. (§1.3.3.24). Retrieved April 12, 2006, from <http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm>

- NIST/SEMATECH. (n.d.). How can we make multiple comparisons? *NLME NIST/SEMATECH e-Handbook of statistical methods* (§7.4.7). Retrieved from <http://www.itl.nist.gov/div898/handbook/prc/section4/prc47.htm>
- NLME: Software for mixed-effects models: <http://cm.bell-labs.com/stat/>
- Nocedal, J., & Wright, S. (1999). *Numerical optimization*. New York: Springer.
- Nonlinear regression in SAS: http://www.ats.ucla.edu/stat/sas/library/SASNLin_os.htm
- Nonprobability sampling: http://www.statcan.ca/english/edu/power/ch13/non_probability/non_probability.htm
- Norcross, J. C., Beutler, L. E., & Levant, R. F. (Eds.). (2005). *Evidence based practice in mental health: Debate and dialogue on the fundamental questions*. Washington, DC: American Psychological Association.
- Normal curve information: http://en.wikipedia.org/wiki/Normal_distribution, <http://mathworld.wolfram.com/NormalDistribution.html>, <http://www.answers.com/topic/normal-distribution>, <http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html>, http://www.tushar-mehta.com/excel/charts/normal_distribution, <http://www-stat.stanford.edu/~naras/jsm/FindProbability.html>
- Nourani, Y., & Andresen, B. (1998). A comparison of simulated annealing cooling strategies. *Journal of Physics A: Mathematical & General*, 31, 8373–8385.
- Nulton, J. D., & Salamon, P. (1988). Statistical mechanics of combinatorial optimization. *Physical Review A*, 37(4), 1351–1356.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). New York: McGraw-Hill.
- The Nuremberg Code. (1947/1996). *British Medical Journal*, 313, 1448.
- O*NET Ability Profiler: www.onetcenter.org
- O'Brien, R. G. (1979). A general ANOVA method for robust test of additive models for variance. *Journal of the American Statistical Association*, 74, 877–880.
- O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, 89(3), 570–574.
- Ockham, W. v. (1982). Quaestiones in III librum sententiarum. In F. E. Kelley & G. I. Etzkorn (Eds.), *Opera theologica: Vol. 6* (pp. 43–97). St. Bonaventure, NY: Franciscan Institute.
- Odeh, R. E., & Fox, M. (1991). *Sample size choice: Charts for experiments with linear models* (2nd ed.). New York: Marcel Dekker.
- OECD handbook for internationally comparative education statistics: Concepts, standards, definitions and classifications. (2004). Paris: OECD.
- Office for Protection from Research Risks, Protection of Human Subjects. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research* (GPO 887-809). Washington, DC: U.S. Government Printing Office.
- Office of Special Education Programs information: <http://www.ed.gov/about/offices/list/osep/index.html?src=mr>
- Ogive (history and definition): <http://www.pballew.net/arithm12.html#ogive>
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241–286.
- Olmstead-Schafer, M., Story, M., & Haughton, B. (1996). Future training needs in public health nutrition: Results of a national Delphi survey. *Journal of the American Dietetic Association*, 96, 282–283.
- Olson, D. H. (1976). *Treating relationships*. Lake Mills, IA: Graphinc.
- Ono, H., Phillips, K. A., & Leneman, M. (1996). Content of an abstract: De jure and de facto. *American Psychologist*, 51(12), 1338–1340.
- Onwuegbuzie, A. J. (1999). Defense or offense? Which is the better predictor of success for professional football teams? *Perceptual and Motor Skills*, 89, 151–159.
- Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools*, 9(1), 73–90.
- Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.
- Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation Press.
- Orme, D. R., Brehm, W., & Ree, M. J. (2001). Armed Forces Qualification Test as a measure of premorbid intelligence. *Military Psychology*, 13(4), 187–197.
- Osborne, A. G., & Russo, C. J. (2006). *Special education and the law: A guide for practitioners* (2nd ed.). Thousand Oaks, CA: Corwin.
- Osgood, C. E. (1957). A behavioristic analysis of perception and language as cognitive phenomena. In J. S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood, & D. Rapaport (Eds.), *Contemporary approaches to cognition: A report of a symposium at the University of Colorado, May 12–14* (pp. 75–118). Cambridge, MA: Harvard University Press.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Ossorio, P. G. (1983). A multicultural psychology. In K. E. Davis & R. M. Bergner (Eds.), *Advances in descriptive psychology* (pp. 13–44). Greenwich, CT: JAI Press.

- Ostrom, T. M. (1989). Interdependence of attitude theory and measurement. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 11–36). Hillsdale, NJ: Erlbaum.
- O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in the ventral temporal cortex. *Journal of Cognitive Neuroscience*, 17, 580–590.
- Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis* (5th ed.). Pacific Grove, CA: Duxbury.
- Owen, L., McNeill, A., & Callum, C. (1998). Trends in smoking during pregnancy in England, 1992–7: Quota sampling surveys. *British Medical Journal*, 317, 728–730.
- Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216–230.
- Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and intelligent laboratory systems*, 58, 261–273.
- Pampel, F. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120–125.
- Pankratz, M., Morrison, A., & Plante, E. (2004). Difference in standard scores of adults on the Peabody Picture Vocabulary Test (rev. 3rd ed.). *Journal of Speech, Language, & Hearing Research*, 47(3), 714–718.
- Parallel coordinate plots software, including CASSATT, a stand-alone Java application: <http://www.rosuda.org/software/>
- Parameter definition: <http://www.animatedsoftware.com/statglos/sgparam.htm>
- Parameters: <http://davidmlane.com/hyperstat/A12328.html>
- Pardy, S. A., Fabrigar, L. R., & Visser, P. S. (2005). Multitrait multimethod analyses. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in the behavioral sciences* (pp. 1343–1347). Chichester, UK: Wiley.
- Parker, R. A., & Berman, N. G. (2003). Sample size: More than calculations. *American Statistician*, 57, 166–170.
- Patel, J. K. (1982). *Handbook of the normal distribution*. New York: Dekker.
- Path diagrams and software for path analysis and SEM: <http://www.spss.com/amos/>
- Patil, K. D. (1975). Cochran's Q test: Exact distribution. *Journal of the American Statistical Association*, 70, 186–189.
- Pavoine, S., Dufour, A. B., & Chessel, D. (2004). From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis. *Journal of Theoretical Biology*, 228, 523–537.
- Pawar, M. (2004). *Data collecting methods and experiences: A guide for social researchers*. Chicago: New Dawn Press.
- Pawar, M. (2004). Learning from data collecting methods and experiences: Moving closer to reality. In M. Pawar, *Data collecting methods and experiences: A guide to social researchers*. Chicago: New Dawn Press.
- Peabody Picture Vocabulary Test authors Lloyd M. Dunn and Leota M. Dunn biographies: <http://www.slpforum.com/bio/dunn.asp>
- Pearson, E. S., D'Agostino, R. B., & Bowman, K. O. (1977). Tests for departures from normality: Comparisons of powers. *Biometrika*, 64, 231–246.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60, 489–498.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157–175.
- Pearson, K. (1909). On a new method of determining the correlation between a measured character A and a character B. *Biometrika*, 7, 96–105.
- Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of USA National Academy of Sciences*, 85(8), 2444–2448. Retrieved from [http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+\[MEDLINE-pmid:'3162770'\]](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+[MEDLINE-pmid:'3162770'])
- Pearson Assessments: <http://www.pearsonassessments.com/index.htm>
- Pearson biography: <http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Pearson.html>
- Pearson r Applet: <http://noppa5.pc.helsinki.fi/koe/corr/cor7.html> (allows you to move the cursor on the scale and thereby change the scatterplot)
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Pena, D., Tiao, G. C., & Tsay, R. S. (2000). *A course in time series analysis*. New York: Wiley.
- Peng, F., Jacobs, R. A., & Tanner, M. A. (1996). Bayesian inference in mixtures of experts and hierarchical mixtures of experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91, 953–960.
- Percival, D. B., and Walden, A. T. (1993). *Spectral analysis for physical applications: Multitaper and conventional univariate techniques*. Cambridge, UK: Cambridge University Press.

- Perfectly random sampling with Markov chains: <http://dbwilson.com/exact/>
- Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237–255.
- Personal Project Analysis, by Brian R. Little: <http://www.brianrlittle.com/ppa/index.htm>
- Personality Research Form: <http://www.sigmaassessment.com/assessments/prf.asp>
- Peskun, P. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60, 607–612.
- Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of Royal Statistical Society A*, 135, 185–206.
- Petraglia, J. (1998). *Reality by design: The rhetoric and technology of authenticity in education*. Mahwah, NJ: Erlbaum.
- Pettersson, M. (2005). The keys to effective IT auditing. *Journal of Corporate Accounting and Finance*, 16(5), 41–46.
- Pfeiffer, C., Windzio, M., & Kleimann, M. (2005). Media use and its impacts on crime perception, sentencing attitudes and crime policy. *European Journal of Criminology*, 2(3), 259–285.
- Phatak, A., & de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, 11, 311–338.
- Pierson, D. J. (2004). How to write an abstract that will be accepted for presentation at a national meeting. *Respiratory Care*, 49(10), 1206–1212.
- Ping, R. A. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166–175.
- Pinho, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.
- Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika*, 59, 239–251.
- Playfair, W. (1786). *The commercial and political atlas*. London: Debrett.
- Plucker, J. A. (Ed.). (2003). *Human intelligence: Historical influences, current controversies, teaching resources*. Retrieved from <http://www.indiana.edu/~intell>
- Poisson, S. D. (1837). *Recherches sur la Probabilite des Jugements en Matiere Criminelle et en Matiere Civile, Precedees des Regles Generales du Calcul des Probabilities*. Paris: Bachelier, Imprimeur-Libraire pour les Mathematiques, la Physique, etc.
- Poisson distribution definitions and illustrations: http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html#poisdistrn
- Poisson distribution generating applet: <http://www.math.csusb.edu/faculty/stanton/probstat/poisson.html>
- Poling, A., Methot, L. L., & LeSage, M. G. (1995). *Fundamentals of behavior analytic research*. New York: Plenum Press.
- Polynomial regression with Stata: <http://www.gseis.ucla.edu/courses/ed230bc1/notes3/curve.html>
- Popper, K. (1965). *The logic of scientific discovery*. New York: Harper & Row.
- Popping, R. (2000). *Computer assisted text analysis*. London: Sage.
- Porter, T. M. (2004). *Karl Pearson: The scientific life in a statistical age*. Princeton, NJ: Princeton University Press.
- Portfolio assessment and implementation checklists resources: <http://www.pampetty.com/assessment.htm>
- Portfolio assessment teacher resources: http://www.phschool.com/professional_development/assessment/portfolio_based_assess.html
- Prather, J. E. (1988). Spurious correlation. *Encyclopedia of statistical science* (Vol. 8, pp. 613–614). New York: Wiley.
- Predictive validity of the Graduate Management Admissions Test report: <http://www.gmac.com/gmac/NewsCenter/Resources/FactSheetGMATValidity.htm>
- Predictive validity of the Graduate Record Examinations advanced psychology test for grade performance in graduate psychology courses: http://www.findarticles.com/p/articles/mi_m0FCR/is_1_36/ai_85007765
- Project Implicit: <https://implicit.harvard.edu/implicit/>
- Propensity score reference list (annotated) with examples of observational studies: <http://sswnt5.sowo.unc.edu/VRC/Lectures/>
- Propp, J., & Wilson, B. (1998). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9, 223–252.
- Proschan, F. (1963). Theoretical explanation of decreasing failure rate. *Technometrics*, 5, 375–383.
- Psycholinguistics definition: <http://www.answers.com/psycholinguistics&r=67> and <http://en.wikipedia.org/wiki/Psycholinguistics>
- The Psychological Corporation: <http://www.psychopr.com>
- Psychology Assessment Resources: <http://www.parinc.com>
- PsycINFO Psychological Abstracts: <http://www.apa.org/psycinfo/>
- Public welfare. *Code of Federal Regulations*, Title 45.
- Pyle, D. (1999). *Data preparation for data mining*. San Francisco: Morgan Kaufmann.
- Pyzdek, T. (2001). *The Six Sigma handbook*. New York: McGraw-Hill.
- Qannari, E. M., Wakeling, I., & MacFie, J. H. (1995). A hierarchy of models for analyzing sensory data. *Food Quality & Preference*, 6, 309–314.
- R Development Core Team. (2003). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- R reference: <http://www.maths.lth.se/help/R/.R/library/SIN/html/fisherz.html>
- R software and manuals: <http://stat.cmu.edu/R/CRAN/>
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.),

- Sociological methodology* (pp. 111–163). Cambridge, MA: Blackwell.
- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.
- Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods*, 7, 504–523.
- Ramsey, P. H., & Ramsey, P. P. (1988). Evaluating the normal approximation to the binomial test. *Journal of Educational Statistics*, 13, 173–182.
- Ramsey, P. H., & Ramsey, P. P. (1990). Critical values for two multiple comparison procedures based on the Studentized range distribution. *Journal of Educational Statistics*, 15, 341–352.
- Ramsey, P. P., & Ramsey, P. H. (1981). Minimum sample sizes for Cochran's test. Retrieved May 9, 2006, from http://www.amstat.org/sections/srms/Proceedings/papers/1981_146.pdf
- Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.
- Random sampling versus randomization—different implications: <http://www.tufts.edu/~gdallal/rand.htm>
- Range and variability: http://psych.rice.edu/online_stat/chapter3/variability.html
- Rangecroft, M. (2003). As easy as pie. *Behaviour and Information Technology*, 22, 421–426.
- Rao, C. (1995). The use of Hellinger distance in graphical displays of contingency table data. In E. M. Tiit, T. Kollo, & H. Niemi (Eds.), *Multivariate statistics and matrices in statistics* (Vol. 3, pp. 143–161). Leiden Netherlands: Brill.
- Rao, C. R. (1973). *Linear statistical inference and applications*. New York: Wiley.
- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21, 24–43.
- Rao, C. R. (1995). Use of Hellinger distance in graphical displays. In E.-M. Tiit, T. Kollo, & H. Niemi (Eds.), *Multivariate statistics and matrices in statistics* (pp. 143–161). Leiden, Netherlands: Brill.
- Rapaport, D., Gill, M., & Schaffer, R. (1968). *Diagnostic psychological testing*. New York: International Universities Press.
- Rasch, G. (1992). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: MESA. (Original work published 1960)
- Rasch Measurement Special Interest Group of the American Educational Research Association: <http://www.raschsig.org>; see also <http://www.rasch.org/rmt/index.htm>
- Rating scales for mood disorders: <http://www.mhsource.com/disorders/diagdepress.html>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Raven Standard Progressive Matrices: <http://www.cps.nova.edu/~cpphelp/RSPM.html>
- Reddon, J. R. (1990). The rejection of the hypothesis of complete independence prior to conducting a factor analysis. *Multivariate Experimental Clinical Research*, 9, 123–129.
- Ree, M. J., & Carretta, T. R. (1994). Factor analysis of ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement*, 54, 457–461.
- Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology*, 44, 321–332.
- Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance; Not much more than g. *Journal of Applied Psychology*, 79, 518–524.
- Reed, J. G., & Baxter, P. M. (1992). Psychological Abstracts and PsycBOOKS. In J. G. Reed & P. M. Baxter, *Library use: A handbook for psychology* (2nd ed., pp. 37–56). Washington, DC: American Psychological Association.
- Regression applet: <http://www.mste.uiuc.edu/activity/regression/#simulation> (allows you to add up to 50 points to a scatterplot and observe applet draw the line of best fit)
- Regression applet: <http://www.stat.wvu.edu/SRS/Modules/Applets/Regression/regression.html> (allows you to enter pairwise data and observe applet compute regression statistics)
- Regression applet: <http://www.stattucino.com/berrie/dsl/regression/regression.html> (allows you to add one or more points to a scatterplot and observe applet draw the regression line and the correlation coefficient)
- Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, 4, 117–126.
- Relationship between alpha, beta, and power: <http://www.sysurvey.com/tips/statistics/type2.htm>
- Reliability Hot Wire: <http://www.weibull.com/hotwire/issue18/relbasics18.htm> (discusses contour bounds and confidence bounds on parameters)
- Rencher, A. C. (1995). *Methods of multivariate analysis*. New York: Wiley.
- Rennie, D. L. (2000). Grounded theory methodology as methodological hermeneutics: Reconciling realism and relativism. *Theory and Psychology*, 10, 481–502.
- Research methods in the social and natural sciences: http://www.mcli.dist.maricopa.edu/proj/res_meth/
- Response to intervention topical links: <http://www.wested.org/nerrc/rti.htm>
- Reynolds, C. R., & Randy, K. W. (2002). *The clinician's guide to the Behavior Assessment System for Children (BASC)*. New York: Guilford.

- Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11(6), 691–714.
- Reynolds, S. L., Saito, Y., & Crimmins, E. M. (2005). The impact of obesity on active life expectancy in older American men and women. *Gerontologist*, 45, 438–444.
- Rhoads, J. C. (2001). Researching authoritarian personality with Q methodology. Part I: Revisiting traditional analysis. *Operant Subjectivity*, 24(2), 68–85.
- Rhoads, J. C. (2001). Researching authoritarian personality with Q methodology. Part II: Intensive study. *Operant Subjectivity*, 24(2), 86–103.
- Rice, J. A. (1994). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.
- Rider, R. A., & Daly, J. (1991). Effects of flexibility training on enhancing spinal mobility in older women. *Journal of Sports Medicine and Physical Fitness*, 31(2), 213–217.
- Rie, F. H. (1966). Depression in childhood: A survey of some pertinent contributions. *Journal of Child Psychology and Psychiatry and Applied Disciplines*, 35(7), 1289–1308.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.
- Ripley, B. D. (2004). *Spatial statistics*. London: Wiley.
- Riverside Publishing: <http://www.riverpub.com/>
- Roach, A. T., & Elliott, S. N. (2004, April). *Alignment analysis and standard setting procedures for alternate assessments*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22, 400–407.
- Robbins, J. M., Webb, D. A., & Sciamanna, C. N. (2004). Cardiovascular comorbidities among public health clinic patients with diabetes: The Urban Diabetics Study. *BMC Public Health*, 5, 15. Retrieved from <http://www.biomedcentral.com/1471-2458/5/15>
- Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The R_V -coefficient. *Applied Statistics*, 25, 257–265.
- Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Erlbaum.
- Roberts Apperception Test for Children product information: www.wpspublish.com
- Robinson, F. (1992). *Love's story told: A life of Henry A. Murray*. Cambridge, MA: Harvard University Press.
- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15–32.
- Rodgers, B. L., & Cowles, K. V. (1993). The qualitative research audit trail: A complex collection of documentation. *Research in Nursing and Health*, 16, 219–226.
- Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah, NJ: Erlbaum.
- Rodriguez, M. C. (2005). Three options are optimal for multiple choice items: A meta analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.
- Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 32, 269–286.
- Rogelberg, S. (Ed.). (2007). *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Rogers, B. (1998). *Pascal: In praise of vanity*. London: Phoenix.
- Rogers, J., Howard, K., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565.
- Rojo, J. (1983). *On Lehmann's general concept of unbiasedness and some of its applications*. Unpublished doctoral dissertation, University of California, Berkeley.
- Rojo, J. (1987). On the admissibility of $c\bar{X} + d$ with respect to the linex loss function. *Commun. Statist. Theory and Meth.*, 16, 3745–3748.
- Ronald A. Fisher biographical essay: http://en.wikipedia.org/wiki/Ronald_Fisher
- Ronald R. Holden's Web site: <http://psyc.queensu.ca/faculty/holden/holden.html>
- Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516–524.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Applied Social Research Methods Series Vol. 6). Newbury Park, CA: Sage.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.
- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons*. New York: Cambridge University Press.
- Rothman, D. J., & Rothman, S. M. (2005). *The Willowbrook wars: Bringing the mentally disabled*

- into the community. New Brunswick, NJ: Aldine Transaction.
- Rotter, J. B. (1966). Generalized expectancies for the internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80, 1–28.
- Rotter, J. B. (1975). Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology*, 43, 56–67.
- Rotton, J., & Cohn, E. G. (2000). Violence is a curvilinear function of temperature in Dallas: A replication. *Journal of Personality & Social Psychology*, 78, 1074–1081.
- Royston, J. P. (1982). Expected normal order statistics (exact and approximate). *Applied Statistics*, 31, 161–165.
- Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115–124.
- Royston, J. P. (1982). The W test of normality. *Applied Statistics*, 31, 176–180.
- Ruane, J. M. (2005). *Essentials of research methods: A guide to social science research*. Malden, MA: Blackwell.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249–264.
- Rucci, A. J., Kirn, S. P., & Quinn, R. T. (1998, January–February). The employee-customer-profit chain at Sears. *Harvard Business Review*, pp. 82–97.
- Ruelle, D. (1988). *Chaotic evolution and strange attractors: The statistical analysis of time series for deterministic nonlinear systems*. Cambridge, UK: Cambridge University Press.
- Rummel, R. J. (n.d.). *Understanding correlation*. Retrieved from <http://www.mega.nu:8080/ampp/rummel/uc.htm>
- Rupp, A. A. (2005). *A framework for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. Manuscript submitted for publication.
- Rupp, A. A. (2006). *The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. Manuscript submitted for publication.
- Russo, C. J., Osborne, A. G., & Borreca, E. (2005). Special education update: The 2004 revisions of the IDEA. *School Business Affairs*, 71(5), 41–44.
- Sacramento City Unified School Dist. Bd. of Educ. v. Rachel H., 14 F.3d 1398 (9th Cir. 1994).
- Salamon, P., Nulton, J., Harland, J. R., Pedersen, J., Ruppeiner, G., & Liao, L. (1988). Simulated annealing with constant thermodynamic speed. *Computer Physics Communications*, 43, 423–438.
- Sales, B. D., & Folkman, S. (Eds.). (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.
- Salk, L. (1973). The role of the heartbeat in the relations between mother and infant. *Scientific American*, 235, 26–29.
- Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics* (2nd ed.). Thousand Oaks, CA: Sage.
- Salkind, N. J. (2006). *Tests and measurement for people who (think they) hate tests and measurements*. Thousand Oaks, CA: Sage.
- Salkind, N. J. (2007). *Statistics for people who (think they) hate statistics: The Excel edition*. Thousand Oaks, CA: Sage.
- Salomon, D. (2004). *Data compression: The complete reference*. New York: Springer.
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman.
- Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (8th ed.). Boston: Houghton Mifflin.
- Sample-size analysis UnifyPow SAS module, by Ralph O'Brien: www.bio.ri.ccf.org/Power/
- Sampling distributions and the central limit theorem: http://people.hofstra.edu/faculty/Stefan_Waner/RealWorld/finitetopic1/sampldistr.html
- Samuda, R. J. (1998). *Advances in cross-cultural assessment*. Thousand Oaks, CA: Sage.
- Sanders, B. (1964). Measuring community health level. *American Journal of Public Health*, 54, 1063–1070.
- Sarafino, E. P. (2005). *Research methods: Using processes and procedures of science to understand behavior*. Upper Saddle River, NJ: Pearson-Prentice Hall.
- Saris, W. E., & Stronkhorst, H. (1984). *Causal modelling in nonexperimental research: An introduction to the LISREL approach*. Amsterdam: Sociometric Research Foundation.
- Sarmonov, O. V. (1966). Generalized normal correlation and two dimensional Frechet classes. *Soviet Doklady*, 168, 596–599.
- Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.
- SAS and SPSS syntax files to assist in exploratory factor analysis: <http://flash.lakeheadu.ca/~boconno2/nfactors.html>
- SAS macro for computing the point, Pearson, and rank biserial coefficients: <http://ftp.sas.com/techsup/download/stat/biserial.html>

- SAS reference: <http://support.sas.com/ctx/samples/index.jsp?sid=494>
- Sattler, J. M. (1992). *Assessment of children: Revised and updated* (3rd ed.). San Diego, CA: Author.
- Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement*, 16, 209–222.
- Savage, I. R. (1953). Bibliography of nonparametric statistics and related topics. *Journal of the American Statistical Association*, 48, 844–898.
- Savickas, M. L. (1984). Career maturity: The construct and its measurement. *Vocational Guidance Quarterly*, 32, 222–231.
- Savickas, M. L. (2000). Assessing career decision making. In E. Watkins & V. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 429–477). Hillsdale, NJ: Erlbaum.
- Sawilowsky, S. (2000). Psychometrics versus datametrics. *Educational and Psychological Measurement*, 60, 157–173.
- Sawilowsky, S. (2000). Reliability. *Educational and Psychological Measurement*, 60, 196–200.
- Sawilowsky, S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development*, 35, 78–88.
- Sawilowsky, S. (2002). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement*, 60, 157–173.
- Sawilowsky, S. (2002). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement*, 60, 196–200.
- Sawilowsky, S. (2003). You think you've got trials? *Journal of Modern Applied Statistical Methods*, 2(1), 218–225.
- Sawilowsky, S. (2004). A conversation with R. Clifford Blair on the occasion of his retirement. *Journal of Modern Applied Statistical Methods*, 3(2), 518–566.
- Sawilowsky, S. (2005). Misconceptions leading to choosing the *t* test over the Wilcoxon Mann-Whitney test for shift in location parameter. *Journal of Modern Applied Statistical Methods*, 4(2), 598–600.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, 111, 353–360.
- Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation via Fortran*. Rochester Hills, MI: JMASM.
- Sayood, K. (2000). *Introduction to data compression* (2nd ed.). San Francisco: Morgan Kaufmann.
- Scanstat program instructions, from J. Hoh & J. Ott: <http://linkage.rockefeller.edu/ott/Scanstat.html>
- Schaefer, E. S., & Bayley, N. (1963). Maternal behavior, child behavior, and their intercorrelations from infancy through adolescence. *Monographs of the Society for Research in Child Development*, 28(3).
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.
- Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing*. New York: Grune & Stratton.
- Scheaffer, R. L., Mendenhall, W., III, & Ott, L. R. (1995). *Elementary survey sampling* (6th ed.). Belmont, CA: Duxbury Press.
- Scheffé, H. (1970). Practical solutions to the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65, 1501–1508.
- Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T. Næs & E. Risvik (Eds.), *Multivariate analysis of data in sensory sciences* (pp. 229–306). New York: Elsevier.
- Schmidt, F. L. (1973). Implications of a measurement problem for expectancy theory research. *Organizational Behavior and Human Decision Processes*, 10, 243–251.
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1, 199–223.
- Schmitt, N., & Stults, D. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 4, 367–373.
- Schoen, R. (1988). *Modeling multigroup populations*. New York: Plenum.
- Scholastic Aptitude Test information: <http://www.collegeboard.com>
- Scholkopf, B., Smola, A., & Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1299–1319.
- Schonlau, M. (2004). Will web surveys ever become part of mainstream research? *Journal of Internet Medical Research*, 6(3) article e31. Retrieved May 14, 2006, from <http://www.jmir.org/2004/3/e31/>
- Schonlau, M., Fricker, R. D., Jr., & Elliott, M. N. (2001). *Conducting research surveys via e-mail and the Web* (chapter 4). Santa Monica, CA: RAND. Retrieved May 14, 2006, from <http://www.rand.org/publications/MR/MR1480/MR1480.ch4.pdf>
- School Bd. of Nassau County v. Arline*, 480 U.S. 273, 287–88 (1987).
- Schrank, F. A., & Flanagan, D. P. (2003). *WJ-III clinical use and interpretation: Scientist-practitioner perspectives*. San Diego, CA: Academic Press.
- Schrank, F. A., Flanagan, D. P., Woodcock, R. W., & Mascolo, J. T. (2002). *Essentials of WJ-III cognitive abilities assessment*. New York: Wiley.
- Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals

- on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, 51(1), 67–78.
- Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage.
- Schroër, G., & Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. *Computational Statistics and Data Analysis*, 20, 185–202.
- Schuurmann, D. (1981). On hypothesis testing to determine if the mean of the normal distribution is contained in a known interval. *Biometrics*, 37, 617.
- Schuler, R. (1995). *Managing human resources* (5th ed.). New York: West.
- Schumacker, R., & Lomax, R. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, CA: Academic Press.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.
- Scott, L. D. (1981). Measuring intelligence with the Goodenough-Harris Drawing Test. *Psychological Bulletin*, 89, 483–505.
- Scree plots: <http://www.rsd-associates.com/mtxscre.htm>
- Seal, H. L. (1967). The historical development of the Gauss linear model. *Biometrika*, 54, 1–24.
- Section 504 of the Rehabilitation Act of 1973, 29 U.S.C.A. § 794(a).
- Sedgwick, J. E. C. (2001). Absolute, attributable, and relative risk in the management of coronary heart disease. *Heart*, 85, 491–492.
- Seidel, S., Walters, J., Kirby, E., Olf, N., & Powell, K. (1998). *Portfolio practices: Thinking through the assessment of children's work*. Washington, DC: National Education Association.
- Semi-interquartile range page: <http://davidmlane.com/hyperstat/A48607.html>
- Seneta, E. (1981). *Non-negative matrices and Markov chains*. New York: Springer.
- Shadish, W. R., & Clark, M. H. (2002). An introduction to propensity scores. *Metodologia de las Ciencias del Comportamiento Journal*, 4(2), 291–298.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shaffer, J. P. (1979). Comparison of means: An *F* test followed by a modified multiple range test. *Journal of Educational Statistics*, 4, 14–23.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–584.
- Shaklee, B. D., Barbour, N. E., Ambrose, R., & Hansford, S. J. (1997). *Designing and using portfolios*. Boston: Allyn & Bacon.
- Shanmugam, R. (1989). Asymptotic homogeneity tests for mean exponential family distribution. *Journal of Statistical Planning and Inference*, 23, 227–241.
- Shanmugam, R. (1991). Incidence rate restricted Poissonness. *Sankhya*, B, 63, 191–201.
- Shanmugam, R. (1992). Factions in Morris' quadratic variance natural exponential family of distributions and a lack of consensus in sample size determination. *Statistics and Probability Letters*, 15, 20–25.
- Shanmugam, R. (1993). Size-biased incidence rate restricted Poissonness and its application in international terrorism. *Applications of Management Science: Public Policy Application of Management Sciences*, 7, 41–49.
- Shanmugam, R. (1995). Goodness of fit test for length-biased data with discussion on prevalence, sensitivity and specificity of length bias. *Journal of Statistical Planning and Inference*, 48, 277–280.
- Shanmugam, R. (2002). A critique of dependence concepts. *Journal of Mathematical Psychology*, 46, 110–114.
- Shanmugam, R. (2003). Index of variation with illustration using molecular data. *Communications in Statistics*, 32, 509–517.
- Shanmugam, R., & Singh, J. (1981). Some bivariate probability models applicable to traffic accidents and fatalities. In C. Tallie, G. P. Patil, & B. A. Baldessari (Eds.), *Statistical distributions in scientific work: Vol. 6* (pp. 95–103). Hingham, MA: Reidel.
- Shanteau, J. (2001). *What does it mean when experts disagree?* Mahwah, NJ: Erlbaum.
- Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York: Guilford.
- Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67, 215–216.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, 52, 591–611.
- Shapiro, S. S., Wilk, M. B., & Chen, H. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63, 1343–1372.
- Shapiro-Wilk test: http://en.wikipedia.org/wiki/Shapiro-Wilk_test
- Shatkey, H. (1995). *The Fourier transform—A primer*. Brown University Technical Report CS-95-37. Retrieved from <http://www.cs.brown.edu/publications/techreports/reports/CS-95-37.html>
- Shaughnessy, J. J., & Zechmeister, E. B. (1994). *Research methods in psychology* (3rd ed.). New York: McGraw-Hill.
- Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Boston: Allyn & Bacon.

- Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan*, 67(1), 57–60.
- Shepard, R. N. (1966). Metric structure in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.
- Shermer, M. (2002). *Why people believe weird things: Pseudoscience, superstition, and other confusions of our time*. New York: Freeman/Owl Book.
- Shih, W. J., & Huang, W.-H. (1992). Evaluating correlation with proper bounds. *Biometrics*, 48, 1207–1213.
- Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford.
- A Short Account of the History of Mathematics* (4th ed., 1908) by W. W. Rouse Ball [excerpt]: http://www.maths.tcd.ie/pub/HistMath/People/Pascal/RouseBall/RB_Pascal.html
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Shrout, P. E., & Fiske, S. T. (Eds.). (1995). *Advances in personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Hillsdale, NJ: Erlbaum.
- Shulkin, B. (2006). *Estimating a population median with a small sample*. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.
- Shumway, R. H., & Stoffer, D. S. (2005). *Time series analysis and its applications*. New York: Springer.
- Šidák, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626–633.
- Sidman, M. (1960). *Tactics of scientific research*. Boston: Publisher's Cooperative.
- Sieber, J. E. (2004). Empirical research on research ethics. *Ethics and Behavior*, 14, 397–412.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Siemer, M., & Joorman, J. (2003). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods*, 8, 497–517.
- Silver, N. C., & Hittner, J. B. (1998). *A guidebook of statistical software for the social and behavioral sciences*. Needham Heights, MA: Allyn & Bacon.
- Silverberg, K. E., Marshall, E. K., & Ellis, G. D. (2001). Measuring job satisfaction of volunteers in public parks and recreation. *Journal of Park and Recreation Administration*, 19(1), 79–92.
- Silverlake, A. C. (1999). *Comprehending test manuals: A guide and workbook*. Los Angeles: Pyrezak.
- Simon, H. (1985). Spurious correlation: A causal interpretation. In H. M. Blalock (Ed.), *Causal models in the social sciences* (2nd ed., pp. 7–21). New York: Aldine.
- Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.
- Simonton, D. K. (1990). *Psychology, science, and history: An introduction to historiometry*. New Haven, CT: Yale University Press.
- Simplification of informed consent documents: <http://www.cancer.gov/clinicaltrials/understanding/simplification-of-informed-consent-docs>
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238–241.
- Simpson, G. G. (1995). *The book of Darwin*. New York: Washington State Press.
- Simpson's paradox: http://exploringdata.cqu.edu.au/sim_par.htm
- Simpson's paradox entry from the *Stanford Encyclopedia of Philosophy*: <http://plato.stanford.edu/entries/paradox-simpson>
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.
- Siperstein, G. N. (1980). *Instruments for measuring children's attitudes toward the handicapped*. Boston: University of Massachusetts.
- Siperstein, G. N., & Gottlieb, J. (1997). Physical stigma and academic performance as factors affecting children's first impressions of handicapped peers. *American Journal of Mental Deficiency*, 81, 455–462.
- Sir Francis Galton: <http://www.mugu.com/galton/>
- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299–321.
- Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement*, 16, 17–31.
- Sirkin, R. M. (Ed.). (2005). *Statistics for the social sciences*. Thousand Oaks, CA: Sage.
- Sivec, H. J., Waehler, C. A., & Panek, P. E. (2004). The Hand Test: Assessing prototypical attitudes and action tendencies. In M. Hersen (Ed.), *Comprehensive handbook of psychological assessment* (Vol. 2, pp. 405–420). New York: Wiley.
- Six Sigma article: http://en.wikipedia.org/wiki/Six_Sigma
- Skarakis-Doyle, E., Miller, L. T., & Reichheld, M. (2000). Construct validity as a foundation of evidence-based practice: The case of the Preschool Language Assessment Instrument. *Journal of Speech-Language Pathology and Audiology*, 24(4), 180–191.
- Skarakis-Doyle, E., Miller, L. T., & Reichheld, M. (2001). Construct validity as a foundation of evidence-based practice: The case of the Preschool Language Assessment Instrument: Erratum. *Journal of Speech-Language Pathology and Audiology*, 25(1), 40.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, 28, 1–39.
- Smerz, J. M. (2005). Cognitive functioning in severe dementia and relationship to need driven behaviors and

- functional status. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 66(3B), 1737.
- Smith, C. D., & Carolyn, D. (1996). *In the field: Readings on the field research experience* (2nd ed.). Westport, CT: Praeger.
- Smith, D. R., Brown, J. A., & Lo, N. C. H. (2004). Applications of adaptive sampling to biological populations. In W. L. Thompson (Ed.), *Sampling rare or elusive species* (pp. 77–122). Washington, DC: Island Press.
- Smith, E., & Smith, R. (Eds.). (2005). *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: JAM.
- Smoothing methods applet for fitting a regression curve to simulated data: <http://www.spss.com/research/wilkinson/Applets/smoothers.html>
- Smouse, P. E. (1981). The Carroll Rating Scale for Depression: II. Factor analyses of the feature profiles. *British Journal of Psychiatry*, 138, 201–204.
- Snider, D. E., & Satcher, D. (1997). Behavioral and social sciences at the Centers for Disease Control and Prevention: Critical disciplines for public health. *American Psychologist*, 52(2), 140–142.
- Sobol, I. M. (1974). *The Monte Carlo method*. Translated and adapted from the second Russian edition by Robert Messer, John Stone, and Peter Fortini. Chicago: University of Chicago Press.
- Society for Research in Child Development Web site: www.srcd.org
- Sociological Abstracts: <http://www.csa.com/factsheets/socioabs-set-c.php>
- Software for estimating time series regression models with autoregressive errors (Output includes regression coefficients, associated bootstrap tests, and a reduced bias autocorrelation estimate.): <http://www.stat.wmich.edu/>, click on Stat Lab→ Software→ Timeseries. A design matrix and outcome data must be entered.
- Southeastern Community College v. Davis*, 442 U.S. 397 (1979).
- Sparrow, S. S., & Cicchetti, D. V. (1989). The Vineland Adaptive Behavior Scales. In C. S. Newmark (Ed.), *Major psychological assessment instruments*, Vol. 2 (pp. 199–231). Needham Heights, MA: Allyn & Bacon.
- Spata, A. V. (2003). *Research methods: Science and diversity*. New York: Wiley.
- Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 205–293.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161–169.
- Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, 3, 271–295.
- Spearman, C. E. (1927). *The abilities of man*. London: Macmillan.
- Spearman rank correlation coefficient: <http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html>
- Spearman's rho applet: http://faculty.vassar.edu/lowry/corr_rank.html (allows you to compute rho and perform a null hypothesis significance test pertaining to this rho value)
- Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 5, 61–77.
- Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurement yearbook*. Lincoln: University of Nebraska Press.
- Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. *Archives of General Psychiatry*, 49, 624–629.
- Spreadsheet functions usage instruction: <http://spreadsheets.about.com/od/excelfunctions/>
- SPSS regression models: http://www.spss.com/regression/data_analysis.htm (contains a downloadable spec sheet)
- SPSS tutorial: How to do a paired samples *t* test: <http://academic.uofs.edu/departments/psych/methods/cannon99/level2c.html>
- Standard deviation computation steps and tutorial: http://davidmlane.com/hyperstat/desc_univ.html
- Standards for Educational and Psychological Testing*: <http://www.apa.org/science/standards.html>
- Stanford-Binet Intelligence Scales: <http://www.riverpub.com/products/sb5/index.html>
- Stanimirova, I., Walczak, B., Massart, D. L., Simeonov, V., Sabyd, C. A., & di Crescenzo, E. (2004). STATIS, a three-way method for data analysis: Application to environmental data. *Chemometrics & Intelligent Laboratory Systems*, 73, 219–233.
- Stanine article: <http://en.wikipedia.org/wiki/Stanine>
- Stankov, L. (2000). The theory of fluid and crystallized intelligence: New findings and recent developments. *Learning and Individual Differences*, 12, 1–3.
- Statistical graphics video lending library: <http://www.amstat-online.org/sections/graphics/library.php>
- Statistical significance discussion: <http://www.statpac.com/surveys/statistical-significance.htm>
- Statistical software package R: <http://www.r-project.org>
- Statistics Canada. (n.d.). *Statistics: Power from data!* Retrieved April 12, 2006, from http://www.statcan.ca/english/edu/power/ch13/non_probability/non_probability.htm
- StatSoft, Inc. (2003). *Electronic textbook*. Retrieved June 14, 2005, from <http://www.statsoftinc.com/textbook/stathome.html>

- StatXact compared to SPSS Exact Tests and SAS software: http://www.cytel.com/Papers/SX_Compare.pdf
- Stecker, P., & Fuchs, L. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*, 128–135.
- Steiger, J. H., & Schonemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). Chicago: University of Chicago Press.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved May 10, 2006, from <http://pareonline.net/getvn.asp?v=9&n=4>
- Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (in press). Using the theory of successful intelligence as a basis for augmenting exams in AP Psychology and Statistics. *Contemporary Educational Psychology*.
- Stephens, T. (1978). *School skills in the classroom*. Columbus, OH: Cedars Press.
- Stephenson, W. (1953). *The study of behavior*. Chicago: University of Chicago Press.
- Stephenson, W. (1977). Factors as operant subjectivity. *Operant Subjectivity, 1*, 3–16.
- Stephenson, W. (1982). Q-methodology, interbehavioral psychology, and quantum theory. *Psychological Record, 32*, 235–248.
- Sternberg, R. J. (1997). *Successful intelligence: How practical and creative intelligence determine success in life*. New York: Plume.
- Sternberg, R. J. (2000). *Handbook of intelligence*. New York: Cambridge University Press.
- Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? *American Psychologist, 52*, 630–641.
- Stevens, J. (1999). Canonical correlations. In J. Stevens (Ed.), *Applied multivariate statistics for the social sciences* (3rd ed., pp. 429–449). Mahwah, NJ: Erlbaum.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
- Stevens, S. S. (1959). *Measurement, psychophysics and utility*. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*. New York: Wiley.
- Stilson, D. W. (1966). *Probability and statistics in psychological research and theory* (pp. 251–255). San Francisco: Holden-Day.
- Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics*. Reading, MA: Addison-Wesley.
- Stockburger, D. W. (2001). *Introductory statistics: Concepts, models, and applications* (2nd ed.). Cincinnati, OH: Atomic Dog.
- Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review, 73*, 142–167.
- Stone, A., & Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16*, 199–202.
- Stone, E. F. (1988). Moderator variables in research: A review and analysis of conceptual and methodological issues. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 6, pp. 191–229). Greenwich, CT: JAI.
- Stone-Romero, E. F. (2002). The relative validity and usefulness of various empirical research designs. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 77–98). Malden, MA: Blackwell.
- Stone-Romero, E. F. (2007). Nonexperimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F. (2007). Quasi-experimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F. (2007). Randomized experimental designs. In S. Rogelberg (Ed.), *The encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
- Stone-Romero, E. F., & Liakhovitski, D. (2002). Strategies for detecting moderator variables: A review of theory and research. *Research in Personnel and Human Resources Management, 21*, 333–372.
- Stone-Romero, E. F., & Rosopa, P. (2004). Inference problems with hierarchical multiple regression-based tests of mediating effects. *Research in Personnel and Human Resources Management, 23*, 249–290.
- Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*, 195–213.
- Strack, S. (1999). *Essentials of Millon inventories assessment*. New York: Wiley.
- Strang, G. (2003). *Introduction to linear algebra*. Wellesley, MA: Wellesley-Cambridge Press.
- Stratified sampling (and other sampling methods) simulation: <http://www.soc.surrey.ac.uk/samp/>
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.
- Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Thousand Oaks, CA: Sage.

- Strauss, H. J., & Zeigler, L. H. (1975). The Delphi technique and its uses in social science research. *Journal of Creative Behavior*, 9, 253–259.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103.
- Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use*. Oxford, UK: Oxford University Press.
- Stronger Accountability component of NCLB: <http://www.ed.gov/nclb/landing.jhtml>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- Structural equation modeling, by David A. Kenny: <http://davidakenny.net/cm/causal.htm>
- Structural equation modeling reference list (journal articles and chapters on structural equation models), by Jason Newsom: <http://www.upa.pdx.edu/IOA/newsom/semrefs.htm>
- Structured Clinical Interview for *DSM-IV* Web page: www.scid4.org
- Strunk, D. J. (1987). *A concise history of mathematics* (4th rev. ed.). New York: Dover.
- Stuart, A. (1955). A test of homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412–416.
- Stuart, A. (1957). The comparison of frequencies in matched samples. *British Journal of Statistical Psychology*, 10, 29–32.
- Student. (1908). On the error of counting a haemocytometer. *Biometrika*, 5, 351–360.
- Student. (1908). The probable error of a mean. *Biometrika*, 6, 1–25.
- Student's *t* Test Applet: <http://nimitz.mcs.kent.edu/~blewis/stat/tTest.html> (allows you to enter data and solve for *t*)
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers*. San Francisco: Jossey-Bass.
- Sullivan, D. F. (1971). A single index of mortality and morbidity. *HSMHA Health Reports*, 86, 347–354.
- Sun, J. (2005). *User readiness to interact with information systems—A human activity perspective*. In ProQuest Digital Dissertations (www.umi.com).
- Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.
- Sunflower plot creation using the R software, including reference index: <http://cran.r-project.org/manuals.html>
- Super, D. E., & Crites, J. O. (1962). *Appraising vocational fitness*. New York: Harper & Brothers.
- Susan Embretson biography: <http://www.psychology.gatech.edu/WhoWeAre/Faculty/bio-SEmbretson.htm>
- Swade, D. (2002). *The difference engine: Charles Babbage and the quest to build the first computer*. New York: Penguin.
- Swaim, K. F., & Morgan, S. B. (2001). Children's attitudes and behavioral intentions toward a peer with autistic behaviors: Does a brief educational intervention have an effect? *Journal of Autism & Developmental Disorders*, 31(2), 195–205.
- Swain, J. J. (2005). Seventh Biennial Survey of Discrete-Event Software Tools. *OR/MS Today*, 32.
- Swayne, D., & Klinke, S. (1999). Introduction to the special issue on interactive graphical data analysis: What is interaction? *Journal of Computational Statistics*, 1, 1–6.
- Sweet, E. M., & Sigman, R. S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York: HarperCollins.
- Taleb, N. N. (2001). *Fooled by randomness: The hidden role of chance in life and in the markets*. New York: Texere.
- Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate* (pp. 36–38). Upper Saddle River, NJ: Prentice Hall.
- Tanner, M. (1996). *Tools for statistical inference*. New York: Springer-Verlag.
- Tate, M. W., & Brown, S. M. (1970). Note on Cochran's *Q* test. *Journal of the American Statistical Association*, 65, 155–160.
- Tate, R. F. (1954). Correlation between a discrete and a continuous variable: Point-biserial correlation. *Annals of Mathematical Statistics*, 25, 603–607.
- Tate, R. F. (1955). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 42, 205–216.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Erlbaum.
- Tavana, M., Kennedy, D. T., & Joglekar, P. (1996). A group decision support framework for consensus ranking of technical manager candidates. *Omega, International Journal of Management Science*, 24, 523–538.
- Tenenhaus, M. (1998). *La régression PLS*. Paris: Technip.
- Ter Braak, C. J. F., & de Jong, S. (1998). The objective function of partial least squares regression. *Journal of Chemometrics*, 12, 41–54.
- Test reviews: <http://www.unl.edu/buros/>
- Test security: <http://www.apa.org/journals/amp/testsecurity.html>
- Testing standards and procedures information: <http://www.apa.org/science/testing.html>
- Tharinger, D. J., & Stark, K. (1990). A qualitative versus quantitative approach to evaluating the Draw-A-Person

- and Kinetic Family Drawing: A study of mood- and anxiety-disordered children. *Psychological Assessment*, 2, 365–375.
- The Data Mine, a data mining-oriented Web site: <http://www.the-data-mine.com/>
- Theory and practice of pseudorandom number generation: <http://random.mat.sbg.ac.at/>
- Thesaurus of Psychological Index Terms*, 10th edition: <http://www.apa.org/books/3100083.html>
- Thibos, L. N. (2003). *Fourier analysis for beginners*. Indiana University School of Optometry Research Library. Retrieved from <http://research.opt.indiana.edu/Library/FourierBook/title.html>
- Thioulouse, J., Simier, M., & Chessel D. (2004). Simultaneous analysis of a sequence of paired ecological tables, *Ecology*, 85, 272–283.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response theory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.
- Thisted, R. A. (1988). *Elements of statistical computing*. New York: Chapman & Hall.
- Thomas, D. B., & McKeown, B. F. (1988). *Q methodology*. Beverly Hills, CA: Sage.
- Thomas Malthus biography with comprehensive links to other sites: <http://cepa.newschool.edu/het/profiles/malthus.htm>
- Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretations*. Beverly Hills, CA: Sage.
- Thompson, B. (1991). Review of generalizability theory: A primer by R. J. Shavelson & N. W. Webb. *Educational and Psychological Measurement*, 51, 1069–1075.
- Thompson, B. (1994). The concept of statistical significance testing. *Practical Assessment, Research & Evaluation*, 4(5). Retrieved November 1, 2005, from <http://PAREonline.net/gettelevision.asp?v=4&n=5>
- Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, 62(2), 157–176.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Society.
- Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement*, 45, 203–209.
- Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050–1059.
- Thompson, S. K. (1992). *Sampling*. New York: Wiley.
- Thompson, S. K., & Seber, G. A. F. (1996). *Adaptive sampling*. New York: Wiley.
- Thorndike, R. L. (1963). *The concepts of over- and under-achievement*. New York: Bureau of Publications, Teachers College, Columbia University.
- Thorndike, R. M. (1990). *A century of ability testing*. Chicago: Riverside.
- Thorndike, R. M. (1994). Correlational procedures in data analysis. In T. Husen & N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 1107–1117). New York: Pergamon.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Thorndike, R. M., & Dinnel, D. L. (2001). *Basic statistics for the behavioral sciences*. Upper Saddle River, NJ: Prentice Hall.
- Thorson, J. A., & Powell, F. C. (1992). Vagaries of college norms for the Edwards Personal Preference Schedule. *Psychological Reports*, 70, 943–946.
- Thun, M. J., Apicella, L. F., & Henley, S. J. (2000). Smoking vs other risk factors as the cause of smoking-attributable deaths. *JAMA*, 284, 706–712.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.
- Thurstone, L. L. (1938). *Primary mental abilities* (Psychometric Monograph No. 1). Chicago: University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.
- Time Series Data Library: <http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/>
- Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*, 35(4), 417–437.
- Timothy W. v. Rochester, N.H., School Dist., 875 F.2d 954 (1st Cir. 1989a), cert. denied 493 U.S. 983 (1989b).
- Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., et al. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children*, 69, 481–494.
- Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.
- Title VI of the Civil Rights Act of 1964, 42 U.S.C.A. §§ 2000 *et seq.*
- Title IX of the Education Amendments of 1972, 20 U.S.C.A. § 1681.
- Togerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Tolor, A., & Brannigan, G. G. (1980). *Research and clinical applications of the Bender-Gestalt Test*. Springfield, IL: Charles C Thomas.
- Tolor, A., & Schulberg, H. C. (1963). *An evaluation of the Bender-Gestalt Test*. Springfield, IL: Charles C Thomas.

- Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.
- Toothaker, L. E. (1993). *Multiple comparison procedures* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07–089). Newbury Park, CA: Sage.
- Torrance, E. P. (1967/1995). Rationale of creativity tests. In E. P. Torrance (Ed.), *Why fly? A philosophy of creativity* (pp. 83–99). Norwood, NJ: Ablex.
- Torrance, E. P., & Goff, K. (2001). *Brief demonstrator Torrance Tests of Creative Thinking*. Bensenville, IL: Scholastic Testing Service.
- Torrance, E. P., & Safter, H. T. (1999). *Making the creative leap beyond*. . . Buffalo, NY: Creative Education Press.
- Towne, R. L. (2003). Review of Illinois Test of Psycholinguistic Abilities—Third Edition. In L. L. Murphy (Ed.), *The fifteenth mental measurements yearbook* (pp. 458–462). Lincoln: University of Nebraska Press.
- Trochim, W. M. K. (2002). *Levels of measurement*. Retrieved from <http://www.socialresearchmethods.net/kb/measlevl.htm>
- Trochim, W. M. K. (2002). *Measurement error*. Retrieved from <https://www.socialresearchmethods.net>
- Trochim, W. M. K. (2002). Quasi-experimental design. In *Research methods knowledge base*. Retrieved June 14, 2005, from <http://www.socialresearchmethods.net/kb/>
- True random number service based on atmospheric noise: <http://www.random.org/>
- True random number service based on radioactive decay: <http://www.fourmilab.ch/hotbits/>
- Tucker, L. R. (1951). *A method for the synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Tukey, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics*, 1, 31–48.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tukey, J. W. (1977). *Exploratory data analysis*. Menlo Park, CA: Addison-Wesley. [See Chapter 1, "Scratching down numbers (stem-and-leaf)"]
- Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of Royal Statistical Society B*, 38, 290–295.
- Turner, S. P. (1979). The concept of face validity. *Quality and Quantity*, 13, 85–90.
- Two Crows: <http://www.twocrows.com/>
- Type II error: <http://www.cmh.edu/stats/definitions/typeII.htm>
- Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1), 140–146.
- Uebersax, J. (2002). *Statistical methods for rater agreement*. Retrieved August 9, 2002, from <http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>
- UMI Dissertation Services. (1997). *Publishing your dissertation: How to prepare your manuscript for publication*. Ann Arbor, MI: UMI.
- UMI ProQuest Digital Dissertations: <http://wwwlib.umi.com/dissertations/>
- Unwin, A. R. (1999). Requirements for interactive graphics software for exploratory data analysis. *Journal of Computational Statistics*, 1, 7–22.
- Urdan, T. C. (2001). *Statistics in plain English*. Mahwah, NJ: Erlbaum.
- U.S. Department of Education. (2005, August). *Alternate achievement standards for students with the most significant cognitive disabilities: Nonregulatory guidance*. Washington, DC: Author. Retrieved from <http://www.ed.gov/policy/elsec/guid/altguidance.pdf>
- U.S. Department of Education (information from the federal Office of Special Education Programs): <http://www.ed.gov/about/offices/list/osep/index.html?src=mr>
- U.S. Department of Education (updates on regulations, articles, and other general information on the Individuals with Disabilities Education Act and special education): <http://www.ed.gov/offices/OSERS/IDEA/>
- U.S. Department of Health and Human Services. Summary of the HIPAA Privacy Rule: <http://www.hhs.gov/ocr/privacysummary.pdf>
- U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services. HIPAA—General Information: <http://www.cms.hhs.gov/hipaa/hipaa2/>
- U.S. Department of Labor. (2002). *Ability Profiler: Administration manual*. Washington, DC: U.S. Government Printing Office.
- U.S. Food and Drug Administration. (1998). *Information sheets: A guide to informed consent*. Retrieved August 27, 2005, from <http://www.fda.gov/oc/ohrt/irbs/informedconsent.html>
- U.S. General Accounting Office. (1999, November). *Survey methodology: An innovative technique for estimating sensitive survey items* (GAO/GGD-00-30). Washington, DC: Author. Retrieved from <http://www.gao.gov/new.items/gg00030.pdf>
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.
- van der Ven, A. H. G. S. (1980). *Introduction to scaling*. Chichester, UK: Wiley.
- Van Rijckevorsel, J., and de Leeuw, J. (Eds.). (1988). *Component and correspondence analysis*. Chichester, UK: Wiley.
- Van Soest, J. (1967). Some experimental results concerning tests of normality. *Statistica Neerlandica*, 21, 91–97.

- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.
- Vapnik, V. (1999). *The nature of statistical theory*. New York: Springer-Verlag.
- Varian, H. R. (1975). A Bayesian approach to real estate assessment. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* (pp. 195–208). Amsterdam: Elsevier North-Holland.
- Vaughn, S., & Fuchs, L. S. (Eds.). (2003). Redefining LD as inadequate response to instruction [Special issue]. *Learning Disabilities Research & Practice*, 18(3).
- Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury Press. (See Chapter 1, “Stem-and-leaf displays”)
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
- Verbeke, G., Lesaffre, E., & Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, 35, 419–434.
- Vernon, P. A. (2000). Recent studies of intelligence and personality using Jackson’s Multidimensional Aptitude Battery and Personality Research Form. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 195–212). Norwell, MA: Kluwer.
- Vineland Adaptive Behavior Scales product information: <http://ags.pearsonassessments.com>
- Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 223–252). Cambridge, UK: Cambridge University Press.
- Visual Statistics with Multimedia, an online tutorial that covers ANCOVA: <http://pages.infinit.net/rlevesqu/spss.htm>
- Vogt, W. P. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.
- von Eye, A., & Schuster, C. (1998). *Regression analysis for the social sciences*. San Diego, CA: Academic Press.
- von Mises, R. (1957). *Probability, statistics, and truth*. New York: Dover.
- Vrij, A. (2005). Criteria-based content analysis—A qualitative review of the first 37 studies. *Psychology, Public Policy, & Law*, 11, 3–41.
- W. Edwards Deming Institute: <http://www.deming.org/>
- Wagner, E. E. (1983). *The Hand Test: Manual for administration, scoring, and interpretation*. Los Angeles: Western Psychological Services. (Originally published in 1962)
- Walach, H., & Schmidt, S. (2005). Repairing Plato’s life boat with Ockham’s razor: The important function of research in anomalies for mainstream science. *Journal of Consciousness Studies*, 12(2), 52–70.
- Wald, A. (1950). *Statistical decision functions*. New York: Wiley.
- Walker, H. M., & McConnell, S. (1988). *Walker-McConnell Scale of Social Competence and School Adjustment*. Austin, TX: PRO-ED.
- Wallenstein, S., & Berger, A. (1981). On the asymptotic power of tests for comparing k correlated proportions. *Journal of the American Statistical Association*, 76, 114–118.
- Walpole, R. E., Myers, R. H., & Myers, S. L. (1998). *Probability and statistics for engineers and scientists* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Walton, M. (1986). *The Deming management method*. New York: Perigee.
- Wanberg, K. W., Horn, J. L., & Foster, F. M. (1977). A differential assessment model for alcoholism. *Journal of Studies on Alcohol*, 38, 512–543.
- Wang, L. (1995). Differential aptitude tests (DAT). *Measurement and Evaluation in Counseling and Development*, 28, 168–170.
- Waugh, R. F. (Ed.). (2005). *Frontiers in educational psychology*. New York: Nova Science.
- Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27, 359–397.
- Web-based software that performs both traditional ANCOVA and a more recently developed, robust ANCOVA (based on the work of McKean & Vidmar, 1994): www.stat.wmich.edu/slab/RGLM/. Click the Online Resources button under the Site Guide, and then click on RGLM.
- Web Center for Social Research Methods, by William Trochim: <http://www.socialresearchmethods.net/>
- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD: Williams & Wilkins.
- Wechsler Adult Intelligence Scale page: http://www.psychcentral.com/psypsy/Wechsler_Adult_Intelligence_Scale
- Wechsler Individual Achievement Test page: <http://www.wiat-ii.com/>
- Wechsler Preschool and Primary Scale of Intelligence page: <http://harcourtassessment.com/haiweb/Cultures/en-US/dotCom/WPPSI-III/WPPSI-III.htm>
- Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, 34(3), 166–176.
- Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. T. (in press). The role of reading comprehension in responses to positively- and negatively-worded items on rating scales. *Evaluation & Research in Education*.
- Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. C. (2003). Profiles of respondents who respond inconsistently to

- positively- and negatively-worded items on rating scales. *Evaluation and Research in Education*, 17(1), 45–60.
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents with different response patterns to positively- and negatively-worded items on rating scales. *Assessment and Evaluation in Higher Education*, 28(6), 587–607.
- Wegener, D. T., & Fabrigar, L. R. (2000). Analysis and design for nonexperimental data: Addressing causal and noncausal hypotheses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412–450). Cambridge, UK: Cambridge University Press.
- Wegman, E. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85, 664–675.
- Wegman, E. J., & Luo, Q. (1997). High dimensional clustering using parallel coordinates and the grand tour. *Computing Science and Statistics*, 28, 361–368.
- Wegmann, P., & Lusebrink, V. B. (2000). Kinetic Family Drawing scoring method for cross-cultural studies. *The Arts in Psychotherapy*, 27, 179–190.
- Weisstein, E. W. (n.d.). *Multivariate normal distribution*. Retrieved from <http://mathworld.wolfram.com/MultivariateNormalDistribution.html>
- Weisstein, E. W. (n.d.). *Quantile-quantile plot*. Retrieved April 12, 2006, from <http://mathworld.wolfram.com/Quantile-QuantilePlot.html>
- Welch, T. A. (1984, June). A technique for high-performance data compression. *IEEE Computer*, 17(6), 8–19.
- Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Newbury Park, CA: Sage.
- Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, 72, 566–575.
- Westlake, W. (1979). Statistical aspects of comparative bioequivalence trials. *Biometrics*, 35, 273–280.
- Wheeler, D. D. (1977). Locus of interference on the Stroop test. *Perceptual and Motor Skills*, 45, 263–266.
- White, A. (1998). Measuring pain. *Acupuncture in Medicine*, 16, 1–11. Retrieved August 6, 2005, from <http://www.medical-acupuncture.co.uk/journal/1998nov/six.shtml>
- Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, 11(4), 522–537.
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford, UK: Oxford University Press.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.
- Wilkinson, L. (1994). Less is more: Two- and three-dimensional graphics for data display. *Behavior Research Methods, Instruments, & Computers*, 26, 172–176.
- Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. (Available for purchase at <http://www.psycinfo.com/psycarticles/1999-03403-008.html>)
- Williams, R. (1982). Comprehensiveness of Psychological Abstracts. *American Psychologist*, 37(3), 339.
- Wilson, D. (2000). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). *Journal of the American Statistical Association*, 92, 49–60.
- Wilson, P. (2004). A preliminary investigation of an early intervention program: Examining the intervention effectiveness of the Bracken Concept Development Program and the Bracken Basic Concept Scale—Revised with Head Start students. *Psychology in the Schools*, 41, 301–311.
- Wimmer, R. D., & Dominick, J. R. (2000). *Mass media research: An introduction*. Boston: Wadsworth.
- Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental designs*. New York: McGraw-Hill.
- Wingograd, P., & Jones, D. L. (1992). The use of portfolios in performance assessment. *New Directions for Educational Reform*, 1(2), 37–50.
- Winkler, W. E. (1994). Advanced methods for record linkage. 1994 *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 467–472). Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/1994_077.pdf
- Winkler, W. E. (2000). Machine learning, information retrieval, and record linkage. 2000 *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 20–29). Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/2000_003.pdf
- Winship, C., & Sobel, M. (2004). Causal inference in sociological studies. In M. Hardy & A. Bryman (Eds.), *Handbook of data analysis* (pp. 481–503). London: Sage.
- Witkin, H. A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, 19, 1–15.

- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.
- Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.
- Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Author.
- Wolk, R. L., & Wolk, R. B. (1971). *The Gerontological Apperception Test*. New York: Behavioral Publications.
- Woo, J., Ho, S. C., & Wong, E. M. C. (2005). Depression is the predominant factor contributing to morale as measured by the Philadelphia Geriatric Morale Scale in elderly Chinese aged 70 years and over. *International Journal of Geriatric Psychiatry*, 20(11), 1052–1059.
- Wood, J. M. (1996). Weighing evidence in sexual abuse evaluations: An introduction to Bayes' theorem. *Child Maltreatment*, 1(1), 25–36.
- Wood, J., Nezworski, M. T., Lilienfeld, S., & Garb, H. (2003). *What's wrong with the Rorschach? Science confronts the controversial inkblot test*. San Francisco: Jossey-Bass.
- Woodcock, R. N. (1998). *Woodcock Reading Mastery Tests Revised/Normative update*. Circle Pines, MN: American Guidance Service.
- Woodcock Reading Mastery Tests Revised research bibliography: <http://www.agsnet.com>
- Wooldridge, J. M. (2005). *Introductory econometrics* (3rd ed.). Belmont, CA: Thomson.
- Worsley, K. J. (1997). An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5, 254–258.
- Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—Again. *NeuroImage*, 2, 173–181.
- Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.
- Xia, X. (2000). *Data analysis in molecular biology and evolution*. Boston: Kluwer Academic.
- Xia, X., & Xie, Z. (2001). DAMBE: Data analysis in molecular biology and evolution. *Journal of Heredity*, 92, 371–373.
- Yalom, I. D. (1980). *Existential psychotherapy*. New York: Basic Books.
- Yi, Q., & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56, 359–378.
- Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: A meta-analytic generalization across studies. *Educational and Psychological Measurement*, 60, 201–223.
- Young, G. R., & Wagner, E. E. (Eds.). (1999). *The Hand Test: Advances in applications and research*. Malabar, FL: Kreiger.
- Youth indicators 1993—School: Outcomes: <http://www.ed.gov/pubs/YouthIndicators/Outcomes.html>
- Zedeck, S. (1971). Problems with the use of “moderator” variables. *Psychological Bulletin*, 76, 295–310.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81, 446–451.
- Zhang, J. (1998). Tests for multiple upper or lower outliers in an exponential sample. *Journal of Applied Statistics*, 25, 245–255.
- Zhang, J., & Yu, K. (2004). The null distribution of the likelihood-ratio test for two upper outliers in a gamma sample. *Journal of Statistical Computation and Simulation*, 74, 461–467.
- Zhang, J., & Yu, K. (2006). The null distribution of the likelihood-ratio test for one or two outliers in a normal sample. *Test*, 15, 141–150.
- Zhang, Z., Sun, L., Zhao, X., & Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33, 61–70.
- Zimmerman, D. W. (2000). Restriction of range and correlation in outlier-prone distributions. *Applied Psychological Measurement*, 24(3), 267–280.
- Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.
- Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *Journal of Experimental Education*, 62, 75–86.
- Ziv, J., & Lempel, A. (1977, May). A universal algorithm for data compression. *IEEE Transactions on Information Theory*, 23(3), 227–343.
- Ziv, J., & Lempel, A. (1978, September). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, IT-24(5), 530–536.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from <http://educ.ubc.ca/faculty/zumbo/DIF/index.html>
- Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505–509). Thousand Oaks, CA: Sage.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

Zytowski, D. G. (1992). Three generations: The continuing evolution of Frederic Kuder's interest inventories. *Journal of Counseling and Development*, 71, 245–248.

Index

- Aalen-Breslow estimator, 3:987
- Abdi, H., 1:284, 2:540–542, 598
- Abduction, grounded theory and, 1:420
- Ability
 - achievement and, 1:8
 - approaches to understanding, 1:2–3
 - definition and dimensions of, 1:2–3
 - See also **Ability tests**
- Ability tests, 1:1–5**
 - assumptions of, 1:3
 - examples of, 1:3–4
 - group differences and, 1:4–5
 - historical overview of, 1:1–2
 - purpose, 1:2
 - technology and, 1:5
- Absolute error loss, 1:237
- Absolute ratio scaling, 1:174
- Absolute zero, 3:825
- Abstracts, 1:5–6**
 - contents of, 1:6
 - descriptive versus informative, 1:5
- Accelerated failure time model, 1:126
- Acceptable quality level (AQL), 1:6
- Acceptance error, 3:1021
- Acceptance limit, 1:7
- Acceptance number, 1:7
- Acceptance phase, 3:897
- Acceptance sampling, 1:6–7**
 - consumer's risk, 1:6–7
 - hypergeometric distribution, 2:444–445
 - producer's risk, 1:6
- Accidental correlation, 3:938
- Achievement
 - ability domain, 1:8
 - definition, 1:2, 8–9
 - intelligence versus, 1:9
 - knowledge/skill domain, 1:8
- Achievement tests, 1:7–11**
 - aptitude tests versus, 1:40, 42–43
 - assessments versus, 1:8
 - classroom, 1:9
 - formative versus summative, 1:9
 - high-stakes, 1:10
 - intelligence tests versus, 1:9, 2:478–479
 - Iowa Test of Basic Skills, 2:487
 - Iowa Tests of Educational Development, 2:488
 - length, 1:8
 - purpose, 1:2, 9
 - reference books for, 1:8
 - reliability, 1:8
 - scoring, 1:7
 - selected versus constructed, 1:7
 - standardized, 1:10
 - validity, 1:9
 - Wechsler Individual Achievement Test, 3:1048
- Acquiescence bias, 2:539, 3:843
- Acta Eruditorum*, 2:500
- Action space, 1:237
- Active life expectancy (ALE), 1:11**
- Activities of daily living (ADLs), 1:11
- Activity dimension of connotative meaning, 3:878, 881
- ACT tests, 1:299
- Ada, Countess of Lovelace, 1:70
- Adams, S. J., 1:240
- Adaptive Behavior Inventory for Children, 3:988
- Adaptive behavior measures
 - Adaptive Behavior Inventory for Children, 3:988
 - Vineland Adaptive Behavior Scales, 3:1044
 - Vineland Social Maturity Scale, 3:1045
- Adaptive sampling design, 1:12**
- Additive conjoint measurement, 2:583–584
- Adjective Checklist (ACL), 1:12–13, 2:767**
- Adjectives, measures using
 - Adjective Checklist, 1:12–13, 2:767
 - semantic differential method, 3:878–882
- Adjusted value, 1:158, 161, 2:649

Admissible probability measurement, 2:460

Affect inventories, 2:641–642

Affective priming, 1:54–55

AFQT. *See* **Armed Forces Qualification Test**

Age

age norms, 1:14–15

basal age, 1:71, 3:951

ceiling age, 3:951

mental age, 2:474, 476

Age norms, 1:14–15

Agglomerative hierarchical algorithms, 1:149

Aggregation of data, 3:892–894

D'Agostino-Pearson K^2 test, 3:885

D'Agostino's Y test, 3:885

AGS Publishing, 1:83

Ahiezer, N. I., 2:567

Akaike, Hirotogu, 1:15

Akaike information criterion (AIC), 1:15–17

application of, 1:16–17

assumptions underlying, 1:16

Bayesian information criterion and, 1:78

definition, 1:15–16

quasi-likelihood information criterion, 1:400

Akritas, M. G., 2:688

Alcohol Use Inventory (AUI), 1:18

Alexander, Leo, 1:322

Algina, J., 1:301

Algina, James, 2:501

Aliases, 1:375, 377, 379

Aliasing, 1:368

Allocation

optimal, 3:969–970

proportional, 3:969–970

sample size, 3:859

stratified random sampling, 3:969–970

Allocation rules

discriminant analysis, 1:267–269

sample-based, 1:268–269

Allport, Gordon, 2:769

Almanac of American Politics, 3:873

Alpha

familywise/experimentwise, 1:104

hypothesis testing, 2:446

inferential statistics, 2:458

meanings of, 1:103–105

significance level, 3:889–891, 966

testwise, 1:104

Type II error, 3:1021

Alternate assessment, 1:18–22

approaches to, 1:20–21

portfolio assessment, 2:776–778

required characteristics, 1:20

research on, 1:22

standards, 1:21

validity, 1:21–22

Alternate forms reliability, 1:155, 2:731, 3:944

Alternating expectation-conditional maximization, 1:313

Alternating least squares method, 2:530

Alternative hypothesis, 1:23–24

directional versus nondirectional, 1:23–24

hypothesis testing, 2:446, 711–712

null hypothesis versus, 1:23

one- and two-tailed tests, 2:712

purpose, 1:23

specification of, 1:24

American Academy of Pediatrics, 1:145

American Council on Education, 1:298

American Cyanamid, 3:1051

American Doctoral Dissertations (ADD), 1:24–25

American Educational Research Association, 1:27, 181,
2:678, 3:947

Joint Committee on Testing Standards, 1:183

American Guidance Service, 2:507, 3:1044,
1045, 1054

American Medical Association, 1:301

American National Standards Institute, 2:429

American Psychiatric Association, 1:82, 121,
145, 147, 148

American Psychiatric Press, 3:977

American Psychological Association (APA), 1:25–26

APS and, 1:49

confidence intervals, 1:178

content validity, 1:183

effect size, 1:301

ethics code, 1:322–325, 2:466

ethics of testing, 1:318–321

evidence-based practice, 1:326

face validity, 1:338

Lightner Witmer Award, 1:417

mission, 1:25

organizational structure, 1:25–26

Personality Assessment Inventory, 2:766

research with human participants, 1:322–325

*Standards for Educational and Psychological
Testing*, 3:947

validity, 3:1034

American Psychological Society. *See* **Association for
Psychological Science**

American Statistical Association (ASA), 1:26–27

annual meetings, 1:27

awards and educational programs, 1:27, 249

Florence Nightingale, 1:413

- formation, 1:27
- publications, 1:27, 2:499–500, 502
- Americans with Disabilities Act (ADA), 1:27–29**
 - California Psychological Inventory, 1:118
 - Sixteen Personality Factor Questionnaire, 3:918
- America Votes*, 3:873
- AMOS, 1:59
- Analysis of covariance (ANCOVA), 1:29–32**
 - ANOVA versus, 1:29–31
 - assumptions and design issues, 1:31–32
 - difference scores versus, 1:259
 - one-factor, 1:31
- Analysis of variance (ANOVA), 1:32–35**
 - ANCOVA versus, 1:29–31
 - assumptions of, 1:34
 - critical values table for, 3:1066–1069
 - differences with repeated measures, 1:35
 - Fisher's LSD, 1:359–360
 - follow-up tests, 1:34
 - grand mean, 1:411–412
 - Kruskal-Wallis one-way, 2:519–520
 - linear models, 1:34
 - moderator variables, 2:624–625
 - O'Brien test for homogeneity of variance, 2:701–704
 - one-factor, 1:30–31
 - sample size, 3:858–859
 - similarities and differences among types, 1:33–34
 - types, 1:32–33
 - See also **One-way analysis of variance**
- Analytical engine, 1:69–70
- Analytical Society, 1:69
- Analytical Theory of Probabilities* (Laplace), 1:129–130
- Analytic Hierarchy Process, 2:727
- Anastasi, Anne, 1:338
- Ancestry approach, in meta-analysis, 2:596
- ANCOVA. See **Analysis of covariance**
- Anderson, D. R., 1:151
- Anderson, T., 2:471
- Anderson, T. W., 2:673
- Animal Behavior*, 2:500
- Animation, 3:903
- Anisotropy, 3:925
- Annals of Mathematical Statistics/Annals of Statistics*, 2:501
- Annals of Probability*, 2:501
- Anonymity, Delphi technique and, 1:241
- ANOVA. See **Analysis of variance**
- Anthropometry, 1:35–39**
 - applications of, 1:37
 - decline of, 1:37
 - definition, 1:35
 - history of, 1:35–36
 - landmarks, 1:36, 37, 38
 - measurement issues, 1:37–38
 - measurement practices, 1:36
 - misapplications, 1:36–37
 - schools of, 1:36
- Anthropometry* (Hrdlička), 1:36
- Applied Psychology Resources, 2:532
- Applied research, 1:39**
 - basic versus, 1:39, 72–73
- Apprehension Bias, 2:430
- Appropriate scoring, 2:651
- Aptitude
 - definition, 1:2, 39
- Aptitude tests, 1:39–43**
 - achievement tests versus, 1:40, 42–43
 - controversy over, 1:41–42
 - definition, 1:2
 - Differential Aptitude Test (DAT), 1:260–261
 - intelligence tests versus, 1:39–40
 - job-specific, 1:40–41
 - multiaptitude battery, 1:41
 - over/underachievement and, 1:42–43
 - purpose, 1:39–40
 - trade tests versus, 1:40
 - utility of, 1:40
- Aquinas, Thomas, 2:707–708
- Area, estimation of, 2:627–628
- Area chart, 1:43**
- Area under the curve (AUC), 1:255–256, 3:1062–1063
- Argument of the complex number, 1:366
- Aristotle, 2:707–708
- Arithmetic coding, 1:227–228
- Arithmetic mean, 1:43–44, 2:586–587**
 - average versus, 1: 65
- Arizona State Lottery, 3:807
- ARMA methodology, 2:633
- Armed Forces Qualification Test (AFQT), 1:44–45**
 - ASVAB and, 1:45, 46
- Armed Services Vocational Aptitude Battery (ASVAB), 1:45–46**
 - AFQT and, 1:45, 46
 - computerized adaptive testing, 1:171
 - occupational use of, 1:41
 - overview of, 1:3–4
 - purposes, 1:45
- Army Alpha test, 1:1
- Army Beta test, 1:1
- Army General Classification Test, 1:45
- Arnold, H. J., 1:48
- Arnold, S. F., 2:688

Ars Conjectandi (Bernoulli), 1:86–87

Artificial neural network (ANN), 1:46–47

Assembly for Scientific and Applied Psychology, 1:49

Assessment

alternate, 1:18–22

definition, 1:8

Assessment of interactions in multiple regression, 1:47–48

Association for Psychological Science (APS), 1:48–49

achievement awards, 1:49

advocacy, 1:49

annual convention, 1:49

formation, 1:48–49

fund for teaching and public understanding, 1:49

publications, 1:49

Student Caucus, 1:49

Association of Research Libraries, 1:24

Assumption, face validity and, 1:339

ASVAB. *See* **Armed Services Vocational Aptitude Battery**

Asymmetry of g , 1:50

Asymptotic relative efficiency (ARE), 1:381, 2:565, 685

Attention Bias, 2:430

Attention deficit disorder (ADD), 1:145–146

Attention deficit/hyperactivity disorder (ADHD), 1:145–146

Attenuation, correction for, 1:50–52

CFA, 1:51

criterion validity, 1:202

history of, 1:51

reliability, 1:51–52

Spearman's method, 1:50–51

Attitudes

definition, 1:53

direct measures, 1:53–54

indirect measures, 1:54–55

physiological measures, 1:55–56

Attitude tests, 1:53–56

Adjective Checklist, 1:12–13

direct, 1:53–54

indirect, 1:54–55

physiological, 1:55–56

semantic differential scales, 3:881

Thurstone scales, 3:1002–1005

Attributable risk, 1:56–57

Attrition bias, 1:57–60

correcting, 1:59–60

detecting, 1:58–59

preventing attrition, 1:58

two effects of, 1:58

See also **Diggle-Kenward Model for Dropout**

Auditory processing, 1:3, 405

Auditory system, data compression for, 1:229

Audit trail, 1:60–61

Authentic assessment. *See* **Performance-based assessment**

Authenticity, 1:61–62

Autocorrelation, 1:62–64

conditions for, 1:62

importance, 1:62

measurement, 1:63–64

relevant, 1:64

Autocorrelation function, 1:64

Autoregressive, integrated, moving averages (ARIMA) modeling, 1:64, 3:1007–1009

Average, 1:65–66

arithmetic mean, 2:586–587

Excel spreadsheet function, 1:326–328, 3:935–937

harmonic mean and, 2:425–426

moving, 2:633–638

Average absolute deviation, 1:66

Average deviation (AD), 1:66–67

Average learning rate, 1:208

Axial coding, 1:419

Babbage, Charles, 1:69–70

Backpropagation algorithm, 1:47

Backpropagation networks, 1:234

Bacon, Roger, 2:707–708

Bagby, R. M., 1:258

Bailey, B. J. R., 1:294

Baker, E. L., 2:759

Baker, R., 2:472

Balakrishnan, N., 3:865

Balla, David, 3:1045

Barbeau, P., 1:38

Barchard, Kimberly A., 1:157

Bar chart, 1:70–71, 412, 2:440, 631

Barclay, Jean, 3:874

Barnett, P. G., 1:65

Baron and Kenny hierarchical multiple regression procedure, 3:997–998

Barycenter, 1:128

Basal age, 1:71, 3:951

Basal score, 1:71

Baseline, 1:258

Baseline hazard function, 3:986–987

Baseline phase, 3:904–905

Base rate, 1:256

- Basic Personality Inventory**, 1:71–72, 2:442
- Basic research**, 1:72–73
- applied versus, 1:39, 72–73
 - hypothesis testing, 1:72
 - purpose, 1:72
- Bates, Marston, 3:1057
- Bayes, Thomas, 1:79
- Bayes factors**, 1:74–76
- alternatives, 1:75
 - Bayesian information criterion and, 1:75, 76–77
 - criticisms of, 1:75
 - utility of, 1:74–75
- Bayesian inferential procedures, 1:75
- Bayesian information criterion (BIC)**, 1:17, 76–79
- alternative formulas, 1:78–79
 - Bayes factor and, 1:75, 76–77
 - definition, 1:76
 - example of, 1:77–78
 - model selection and, 1:78
- Bayesian statistics**, 1:79–81
- bivariate distributions, 1:99
 - decision theory and, 1:239–249
 - multiple imputation for missing data, 2:664
 - posterior distribution, 2:781
 - prior distribution, 2:786–787
- Bayes' theorem, 1:79–80, 255
- Bayley Scales of Infant and Toddler Development (Bayley-III), 1:81
- Bayley Scales of Infant Development (BSID-II)**, 1:81–82
- Beck, Aaron T., 1:82
- Beck Depression Inventory (BDI)**, 1:82–83, 121, 2:605, 769, 3:1050
- Becker, R. A., 1:415
- Bedrick, Edward J., 1:96, 97
- Begg, Colin, 2:597
- Begging the question, 1:257
- Behavioral Publications, 1:403
- Behavior Assessment System for Children (BASC-2)**, 1:83–84, 3:846
- Behrens-Fisher Test**, 1:84–85
- Belanger, A., 3:804
- Bell, Alexander Graham, 1:27
- Bell curve, 1:206, 2:690–695
- The Bell Curve* (Herrnstein & Murray), 1:45
- Belmont Report, 1:322, 323, 2:466, 481
- Bender, Lauretta, 1:85
- Bender Visual Motor Gestalt Test**, 1:85–86, 3:988
- Beneficence, in research, 1:323–324
- Bentler, Peter, 2:501
- Benzécri, J. P., 2:653
- Berger, A., 1:154
- Berger, Vance W., 2:501
- Berk, Richard, 3:831
- Bernoulli, Jakob**, 1:86–87, 2:772
- Bernoulli, Johann, 1:86
- Bernoulli distribution, 2:619, 622
- Bernoulli trials, 1:87
- Bertin, Jean-Jacques, 1:414
- Best Linear Unbiased Predictors (BLUPs), 2:615
- Beta error, 3:1021
- Beta weights, 2:549–550
- Between-groups ANOVA, 1:33, 2:713
- Between-study cosine matrix, 1:286, 3:957–958
- Bias
- acquiescence, 2:539
 - attrition, 1:57–60
 - Delphi technique and, 1:243
 - Hello-Goodbye Effect, 2:429–430
 - hidden, 2:706
 - item and test, 2:489–493
 - naive estimates and, 2:705
 - order effects, 3:810
 - overt, 2:705–706
 - prediction, 2:492
 - publication, 1:353, 2:596, 597
 - selection, 2:689
 - systematic measurement error, 2:585
- Big Five personality domains, 2:679, 769–770, 3:918
- Bihistogram, 2:440
- Bimodal data set, 2:586
- Bimodal distribution, 2:623
- Binet, Alfred, 1:1, 39, 2:473–474, 476–477, 3:950
- Binet-Simon Intelligence Scale, 1:1, 39
- Binomial distribution/binomial and sign tests**, 1:87–89
- Binomial effect size display (BESD), 1:303
- Binomial test**, 1:88, 89–90
- Binomial theorem, 1:87
- Bins, 2:623, 651
- See also Class interval*
- Bioinformatics**, 1:90–94
- disciplines contributing to, 1:90
 - methodologies, 1:91
 - questions addressed by, 1:91
 - software programs, 1:92–93
 - subtopics, 1:92
 - utility of, 1:91
- Biometrics Bulletin*, 1:27
- Biometrika*, 2:750

- Biostatistics, 1:314
 Birnbaum, A. S., 2:616
Biserial correlation coefficients, 1:94–97
 alternative estimators, 1:96–97
 Pearson's biserial correlation, 1:96
 point biserial correlation, 1:94–96
Bivariate distributions, 1:97–102
 concordant versus discordant, 1:102
 selective continuous probability, 1:100–101
 Bivariate logistic distribution, 1:102
 Blair, R. Clifford, 2:501
 Blalock, H. M., 3:397
 BLAST (Basic Local Alignment Search Tool), 1:92–93
 Blin (physician), 2:473
 Blitstein, J. L., 2:616
 Block designs. *See* **Fractional randomized block design**;
 Randomized block *F* test; Randomized complete
 block designs
Board of Education of the Hendrick Hudson Central
School District v. Rowley, 2:455–456
 Boik, Robert, 2:501
 Bolger, N., 3:996, 998
 Boltzmann distribution, 2:570–571, 3:896–897
 Bona fide pipeline, 1:54–55
 Bond, T., 2:583–584
Bonferroni, Carlo Emilio, 1:103, 294
 Bonferroni's inequalities, 1:294
Bonferroni test, 1:35, 103–106, 2:646
 Boodoo, G., 2:758
 Boole, George, 1:294
 Boole approximation, 1:105
 Boole's inequality, 1:294
 Boosting, 1:270
 Boothby, J. L., 1:83
 Bootstrap method, 1:169, 2:688
 Bornstein, M. H., 1:168
 Borrello, G. M., 2:549
 Bound, J., 2:471
 Bowker, Albert, 1:107, 2:580
Bowker procedure, 1:107–111
 a posteriori comparisons, 1:108
 application, 1:108–109
 significance testing, 1:107–108
 Box counting method, 1:374
 Box-Pierce test, 1: 64
Box plot (box and whisker plot), 1:111–113,
 413, 414
 Boyle, Robert, 1:86
Bracken Basic Concept Scale—Revised, 1:113–114
 Brainerd, Walt, 2:501
 Brazelton, T. Berry, 2:680
 Breckenridge, B. G., 2:554
 Breiman, L., 1:270
 Brewer, E. M., 1:212–213
 Briggs, Katharine, 2:675
 Broadbooks, W. J., 3:850–851
 Broca, Paul, 1:36
 Brogden, Hubert, 1:96
 Browder, D. M., 1:20
 Brown, S. W., 1:152
 Brownian motion, 2:568
Bruno, James Edward, 1:114–115, 2:460–463
 Brunswick, Duke of, 1:385
 Brushing, 1:415–416
 Bryant, P. G., 1:329–331
 Buffon, George Louis Leclerc, Comte de, 2:626, 628
 Buffon Needle problem, 1:92
Bulletin des Sciences Mathématiques, 2:501
 Burnham, K. P., 1:151
 Burnight, K. P., 1:187
 Burns, Robert C., 2:511
 Buros, Luella Gubrud, 1:115
 Buros, Oscar K., 1:115, 3:846
 Buros Center for Testing, 1:115
 Buros Institute for Assessment Consultation and
 Outreach, 1:115–116
Buros Institute of Mental Measurements, 1:115–116,
 2:478, 488, 3:846
 Burt, C., 3:849
 Burt matrix, 2:653
 Burt table, 1:192
 Busch, Douglas, 1:297
 Bush, George W., 2:437
 Business effectiveness, 1:249

 Calculus, 1:86
 California Personality Inventory (CPI), 2:770
California Psychological Inventory (CPI),
 1:117–118
 Cambridge Scientific Abstracts (CSA), 3:923
 Campbell, D., 1:179, 2:666, 3:805, 807
 Canonical correlation analysis, 1:265, 3:1037–1041
 Canons, 1:35
Career Assessment Inventory (CAI), 1:118
 Career choice readiness, 1:120
Career Development Inventory, 1:119
 Career instruments. *See* Vocational instruments
 Career inventories, content- versus process-oriented,
 1:119, 120
Career Maturity Inventory (CMI), 1:120–121

- Carnegie, Andrew, **1:27**
Carnegie Foundation for the Advancement of Teaching, **1:298**
Carroll, Bernard, **1:121**
Carroll, John, **1:404, 2:477–478**
Carroll Depression Scale (CDS), 1:121
Case control studies, **3:806, 833**
Casella, G., **1:85**
Casella, R. L., **1:85**
Castañeda, M. B., **1:35**
Catalan, Eugène, **2:500**
Categorical data
 area charts, **1:43**
 bar charts, **1:70**
 chi-square tests, **1:133**
 pie charts, **2:770**
 Simpson's paradox, **3:892–894**
 See also **Categorical variables**
Categorical validity, **1:254**
Categorical variables, 1:122, 2:554
Category scaling, **3:864**
Cattell, James McKeen, **2:441**
Cattell, Raymond B., **1:204, 404, 2:441, 477–478, 769, 3:917–918, 972**
Cattell-Horn-Carroll (CHC) theory of cognitive abilities, **1:2–3, 404, 2:477–478, 507, 3:1054**
Cattell-Horn Gf-Gc theory of intelligence, **1:404, 3:1053–1054**
Cauchy-Schwartz inequality, **1:160**
Causal analysis, 1:123–125
 basic concepts, **1:123–124**
 benefits, **1:123**
 bivariate distributions, **1:97**
 correlation and, **1:161, 3:937–939**
 criteria of causality, **1:124**
 generative causation, **3:938–939**
 mediating effects, **3:995–998**
 path analysis, **2:745–748**
 prediction and, **1:123**
 quasi-experiments, **3:805**
 regularity theory, **3:937**
 statistics and, **1:123**
 structural equation modeling, **1:124–125, 3:973–974, 976**
Causal modeling, **3:973, 997**
Cefai, C., **1:112**
Ceiling age, **3:951**
Ceiling effect, **2:702–704**
Ceiling scores, **1:71**
Cells, **3:812**
Censored data, 1:125–126, 3:982–983
 See also **Missing data method**
Censored variables, **3:831**
Census databases, **3:873**
Center for Epidemiological Studies of Depression Scale, **1:121**
Center for Statistics Education, **1:27**
Centering
 Generalized Procrustes Analysis, **1:395**
 grand mean, **2:435**
 groups mean, **2:435**
 hierarchical linear modeling, **2:435**
Center of gravity, **1:128**
Center of mass, **1:128**
Centers for Disease Control and Prevention (CDC), 1:126–127
Central Limit Theorem, 1:127–128
 Delta method, **1:247**
 normal curve, **2:690**
Central tendency, **1:243**
 See also **Average; Mean; Measures of central tendency; Median**
Centroid, 1:128–129
 examples of, **1:128–129**
 notations and definition, **1:128**
 properties, **1:129**
Cephalic Index, **1:36**
Cephalometry, **1:35**
CFA. *See* Confirmatory factor analysis
Champion, **3:914–915**
Chance, 1:129–133
 as artifact, **1:131–132**
 definition, **1:129**
 hypothesis testing and, **2:445–446**
 imprecision of, **1:130**
 loglinear analysis, **2:555**
 null hypothesis and, **2:695**
 Poisson distribution, **2:772–775**
 probability and, **1:129–130**
 statistical significance and, **3:963, 965–966**
 technical use of, **1:131**
 unmeasured determinants and, **1:130–131**
Chang, H., **1:172**
Change score, **1:258**
Characteristic equation, **1:304, 309**
Characteristic vectors, **1:304**
Charmaz, Kathy, **1:418**
Charts
 area, **1:43**
 bar, **1:70–71, 412, 2:440, 631**

- line, 2:544
- pie, 1:412, 2:770
- Chasles, Michel, 2:772
- Chauncey, Henry, 1:298
- CHC theory. *See* Cattell-Horn-Carroll (CHC) theory of cognitive abilities
- Chebyshev, Pafnuty Lvovich, 2:568
- Child Behavior Checklist, 3:921
- Child Development*, 3:922
- Child development, 3:922–923
- Child Development Abstracts and Bibliography*, 3:922
- Children's Academic Intrinsic Motivation Inventory** (CAIMI), 1:138–139
- Chi-square test, critical values for, 3:1071
- Chi-square test for goodness of fit**, 1:133–135, 407–408
 - assumptions underlying, 1:134
 - example of, 1:134
 - Pearson and, 2:750
 - Shapiro-Wilk versus, 3:885
 - SPSS calculation of, 1:135
 - structural equation modeling, 3:975
 - utility of, 1:134
- Chi-square test for independence**, 1:135–138
 - assumptions and accuracy, 1:136–137
 - calculations, 1:135–136
 - Fisher exact probability test versus, 1:137
 - larger designs, 1:138
 - SAS JMP computation of, 1:137–138
- Christensen, Anne-Lise, 2:561
- Christianity, 2:707–708
- Cicchetti, Domenic, 3:1045
- Cincinnati Bell, 3:807
- Circularity, 3:836
- City-block distance, 1:283
- Civil Rights Act of 1964, Title VI, 3:876
- Class boundaries. *See* **Class interval**
- Classical ratio estimate (CRE), 2:431–432
- Classical reliability theory. *See* **Classical test theory**
- Classical test theory** (CTT), 1:140–143
 - alternatives to parallel model, 1:142
 - assumptions of, 1:141–142
 - brief history, 1:142
 - controversies over, 1:142
 - impact of, 1:142–143
 - true score and, 3:1014–1016
- Classification
 - aptitude tests for, 1:41
 - discriminant analysis, 1:267–270
 - support vector machines, 3:978–980
- Classification accuracy studies, 1:201–202
- Classification and regression tree** (CART), 1:143–145
 - benefits, 1:143
 - discriminant analysis, 1:270
 - example of, 1:144–145
 - software packages, 1:143
- Class interval**, 1:139–140
- Class limits. *See* **Class interval**
- Classroom achievement tests, 1:9
- Clemans, William, 1:97
- Clemans-Lord estimator, 1:97
- Clerical performance selection instruments, 2:607
- Cleveland, W. S., 1:415
- Clinical Assessment of Attention Deficit** (CAT), 1:145–146
- Clinical Assessment of Behavior** (CAB), 1:147–148
- Clinical Assessment of Depression** (CAD), 1:148–149
- Clinical Assessment Scales for the Elderly, 3:846
- Clinical Global Rating of Depression, 1:121
- Clinton, Bill, 2:426
- Cloning, 1:93
- Closed-ended questions
 - essays, 1:315
 - questionnaires, 3:809
 - self-report measures, 3:877
- Closed-form Newton-Cotes integration methods, 3:895
- Cluster analysis**, 1:149–150
 - clustering algorithms, 1:149
 - hierarchical, 2:503
 - k-means, 1:235, 2:503–506
 - nonhierarchical, 2:503
 - scan statistic, 3:865–867
- Clustering algorithms, 1:149, 235
- Cluster sampling**, 1:150–151
- Coarse factor scores, 1:347
- Cochran, William G., 1:151, 2:500
- Cochran Q test**, 1:151–155
 - a posteriori comparisons, 1:153
 - application of, 1:153–154
 - development, 1:152–153
 - motivation, 1:152
- Code Excited Linear Predictive (CELP) algorithm, 1:228
- Code of Federal Regulations, Protection of Human Subjects, 2:481
- Code of Professional Responsibilities in Educational Measurement, 2:489
- Code on Dental Procedures and Nomenclature, 2:429
- Coder reliability, 2:596
- Coding
 - data, 1:419
 - reverse coded items, 3:844–845

- Coefficient alpha, 1:155–158**
 Cronbach and, 1:204
 formula, 1:156
 hand calculation of, 1:156
 inferential procedures, 1:157
 interpretation, 1:157
 other names for, 1:155
 reliability, 1:155–157
 reliability estimates, 2:519
 SPSS calculation of, 1:156–157
 uses, 1:155–156
- Coefficient of determination, 3:1030
- Coefficient of equivalence, 2:731
- Coefficient of reproducibility, 1:423
- Coefficient of variation (CV), 2:774
- Coefficients of correlation, alienation, and determination, 1:158–161**
 causation and correlation, 1:161
 example of, 1:159
 geometric interpretation, 1:160–161
 interpretation, 1:160–161
 linear and nonlinear relationship, 1:160
 notations and definition, 1:158–159
 outliers, 1:160
 population correlation estimation, 1:161
 properties, 1:159–160
 significance testing of r , 1:161
See also **Correlation coefficient**
- Cognitive Abilities Test (CogAT), 1:162**
- Cognitive psychometric assessment (CPA), 1:163**
 Cognitive Abilities Test, 1:162
 Embedded Figures Test, 1:313–314
 Gf-Gc theory of intelligence, 1:404
- Cognitive psychometric models, 1:163
- Cogswell, William, 1:27
- Cohen, J., 3:859
- Cohen, Jacob, 1:48, 2:596
- Cohen's d , 2:596–597
- Cohen's kappa, 1:164–166**
- Collapsing of data, 3:892–894
- College Board, 1:411
- College Entrance Examination Board, 1:298
- Collinearity, 2:548
- Commercial databases, 3:873
- Committee on Child Development, 3:922
- Common correlation, 1:401
- Common factor model, 1:341–342, 346
- Common factors, 1:332, 333–334, 341, 345
- Comparative fit index, 3:975
- Comparisons
 multiple, 2:644–647
 planned, 2:646
 post hoc, 2:646
- Compensatory model, 1:163
- Complementary events, 1:104
- Complete independence hypothesis, 1:166–167**
- Complete mediation, 3:995
- Complete sufficient statistic, 2:773
- Completion items, 1:167–168**
- Components, 1:211
- Compound symmetry, 1:401, 3:836
- Comprehensive System, 3:848
- Compromise, 1:286–289, 3:955, 957–962
- Computational statistics, 1:168–170**
 data mining/knowledge discovery, 1:170
 meanings of, 1:168
 Monte Carlo methods, 1:170
 resampling and data-partitioning methods, 1:169
 software packages, 1:170
 statistical computing, 1:168–169
- Computerized adaptive testing (CAT), 1:170–173**
 advantages, 1:171
 future issues, 1:173
 goal, 1:171
 item selection, 1:171–173
 large-scale, 1:171
 nonstatistical constraints, 1:172
 paper-and-pencil versus, 1:173
 Rasch measurement, 3:824–825
- Computers
 Babbage, 1:69–70
 simulation experiments, 3:901–903
 See also Software
- Comrey, Andrew L., 1:174**
- Comrey Personality Scales (CPS), 1:174, 175**
- Conant, James, 1:298
- Conceptual maturity, 1:406
- Concordant bivariate, 1:102
- Concourse, 3:799
- Concurrent validity, 1:200–201, 3:1029, 1033
- Conditional logistic regression, 3:833
- Conditional probability, 1:175–176**
- Confidence-based learning, 2:460
- Confidence intervals (CIs), 1:177–178**
 Fisher's Z transformation, 1:362
 narrow, importance of, 3:945
 random sampling, 3:818
 regression analysis, 1:214
 sample size, 3:856–857
 statistical decision theory and, 1:237, 238–240

- Confidence levels, **1:177**
- Confidentiality, **1:319, 2:455, 466**
- Confirmatory analysis, exploratory data analysis versus, **1:328**
- Confirmatory factor analysis (CFA)
 - construct validity, **1:178–179**
 - correction for attenuation, **1:51**
 - exploratory versus, **1:332, 344**
 - model evaluation, **1:343–344**
 - model fitting, **1:343**
 - model modification, **1:344**
 - model specification, **1:343**
 - multitrait multimethod matrix, **2:668–669**
 - overview of, **1:343–344**
 - parameter invariance, **2:733**
 - structural equation modeling and, **3:973**
- Confounding, **3:1037**
- Confounding contrast, **1:376–377**
- Congeneric model, **1:142**
- Conger, A. J., **3:981**
- Congruence coefficient. *See* **R_v and congruence coefficients**
- Conjoint additivity, **2:583**
- Conjoint measurement, **2:583**
- Conjunctive model, **1:163**
- Connotative meaning, **3:878**
- Conover, W. J., **3:884**
- Conover, William J., **2:501, 513**
- ConQuest, **3:822**
- Consensus
 - Delphi technique, **1:244**
 - interrater reliability, **2:484**
- Consensus configuration, **1:394–396**
- Consensus proportion, **1:394, 396**
- Consensus theory, **1:241**
- Consequences-of-testing validity, **3:949, 1034**
- Consolidated Omnibus Budget Reconciliation Act (1985), **2:426**
- Constancy of error variance assumption, **2:546**
- Constant, **2:549**
- Constant thermodynamic speed cooling schedule, **3:899**
- Construct, definition of, **3:948–949**
- Constructed-response items, **1:7, 167**
- Construct method pairings, **2:666–668**
- Construct validity, 1:178–180**
 - differential effects, **1:178**
 - factor analysis, **1:178**
 - history of, **3:1034**
 - Jackson and, **2:497**
 - multitrait multimethod matrix, **1:179, 2:666–669**
 - Pearson product-moment correlation, **1:178**
 - quasi-experiments, **3:806**
 - Standards for Educational and Psychological Testing* on, **3:949**
 - weak versus strong, **3:1034**
- Consulting Psychologists Press, **1:117, 3:971**
- Content analysis, **3:999**
- Content-balancing constraints, **1:172–173**
- Content representativeness, **1:183**
- Content validity, 1:181–183**
 - appropriateness of test development process, **1:182**
 - conducting study of, **1:182–183**
 - definition, **1:181**
 - domain definition, **1:181**
 - domain relevance, **1:182**
 - domain representation, **1:181–182**
 - history of, **1:183, 3:1034**
 - Standards for Educational and Psychological Testing* on, **3:949**
- Contingency tables
 - chi-square test for independence, **1:135**
 - mosaic plots, **2:631**
- Continuous data, **2:586**
- Continuous Fourier transform, **1:365, 366**
- Continuous models, **3:901**
- Continuous updating estimator, **1:391**
- Continuous variable, 1:183–185**
 - definition, **1:184**
 - discrete versus, **1:184–185**
- Contour plot, 1:185–186**
- Control
 - regression and, **3:831–832**
 - statistical, **3:913, 917**
- Convenience sampling, 1:186–188**
 - cluster sampling versus, **1:151**
 - nonprobability sampling, **2:689**
- Convergent thinking, **2:477**
- Convergent validity, **2:666–668**
- Convolution, **1:366**
- Conway, J. M., **1:335**
- Cook, Thomas D., **3:805**
- Cooling schedules, **3:898–899**
- Cooper, Harris, **2:595**
- Coping processes, **1:188**
- Coping Resources Inventory for Stress (CRIS), 1:188**
- Copyright, **1:321**
- Corbin, Juliet, **1:418–419**
- Correct rejection, **2:697–698**
- Correlation
 - accidental, **3:938**

- causation versus, 3:937
- covariance and, 1:195
- direct, 3:938
- Galton and, 1:383
- genuine, 3:938
- illusory, 3:938
- indirect, 3:938
- Kendall rank, 2:508–510
- nonsense, 3:938
- Pearson and, 2:750
- profile analysis, 2:788
- spurious, 3:937
- types, 3:938
- Correlational designs, 3:806
- Correlation coefficients, 1:189–190**
 - homogeneity of, 2:626
 - multiple, 2:648–651
 - Spearman's rho, 3:927–933
 - values for rejecting null hypothesis, table of, 3:1070
- See also **Coefficients of correlation, alienation, and determination; Multiple correlation coefficient; Part and partial correlation**
- Correlogram, 1: 64
- Correspondence analysis, 1:191–193**
 - dimension reduction, 1:265
 - discriminant, 1:270–275
 - joint, 2:656
 - multiple, 1:192–193, 2:651–656
 - objective, 1:191
 - simple, 1:191–192
- Costa, Paul, 2:679
- Cost-benefit considerations, ethics and, 1:323, 2:481
- Counterfactual approach, 1:124–125
- Coupling From the Past, 2:573
- Covariance, 1:194–195, 2:741**
- Covariance components modeling. See **Hierarchical linear modeling**
- Covariance-level interdependence, 2:489
- Covariance structure analysis, 3:973
- Covariates, ANCOVA and, 1:29–32
- Cox & Snell R Square, 2:554
- Cox proportional hazard model, 1:126, 3:831, 986–987
- Cramer-von Mises test, 3:885
- Craniofacial measurements, 1:35, 37
- Creativity
 - Thinking Creatively in Action and Movement, 3:1011–1012
 - Torrance and, 3:1009–1010
 - Torrance Tests of Creative Thinking, 3:1010–1011
- Crick, Francis H. C., 1:90
- Criminals, anthropometry and, 1:37
- Criteria-based content analysis, 1:388
- Criterion-referenced interpretations, 1:9
- Criterion-referenced tests (CRTs), 1:195–200**
 - administration of test, 1:198
 - construction, 1:197–199
 - definition, 1:195, 197
 - domain-referenced versus objective-referenced tests, 1:197
 - interpretation, 1:198
 - items, 1:198–199
 - naming the test, 1:198
 - norm-referenced versus, 1:195
 - objectives represented by, 1:198
 - scoring, 1:198
 - standard error of measurement (SEM), 1:199
 - taxonomies of objectives, 1:197
 - validity, 1:200
- Criterion validity, 1:200–202**
 - classification accuracy, 1:201–202
 - concurrent versus predictive validity, 1:200–201
 - history of, 3:1033
 - projective testing, 2:769
 - Standards for Educational and Psychological Testing* on, 3:949
 - validity coefficient, 3:1029–1032
- Criterion variable, 1:200, 249, 2:545
- Crites, John O., 1:120
- Critical differences, 2:681, 3:1016–1018
- Critical ratio, 1:360
- Critical value, 1:202–203, 2:458–459**
- Critical values tables
 - analysis of variance/*F* test, 3:1066–1069
 - areas beneath normal curve, 3:1062–1063
 - chi-square test, 3:1071
 - correlation coefficient values for rejection of null hypothesis, 3:1070
 - t* values for rejection of null hypothesis, 3:1064–1065
- Cronbach, Lee J., 1:155, 157, 203–204, 338, 2:430, 3:1034**
- Cronbach's alpha, 1:155, 203, 2:519, 3:835
- Cross categorization, 1:133
- Cross-classified categorical data, 3:892–894
- Cross Industry Standard Process for Data Mining (CRISP-DM), 1:233
- Cross-modality matching, 3:863
- Cross-product matrices, 2:599–601
- Cross-validation methods, 1:169

Crystallized intelligence (Gc)
 definition, **1:3**, 405
 Stanford-Binet Intelligence Scales, **3:951**
 theory of, **1:404**, **2:477**
Culling, **1:234**
Cultural bias, **1:204–205**, **2:430**, 478, 757, **3:1043**
Culture Fair Intelligence Test (CFIT), **1:204–205**
Cumming, G., **1:26**
Cummings, Blanche, **1:203**
Cumulative distribution function (CDF),
 2:691–692, **3:982**
Cumulative frequency distribution, **1:205–206**
Cumulative hazard function, **3:982**
Cureton, Edward, **1:97**
Current, R. R., **1:316**
Current Directions in Psychological Science, **1:49**
Current Index for Statistics, **1:27**
Current Procedural Terminology, **2:429**
Curriculum, narrowing of, **2:438**
Curriculum-based measurement (CBM),
 1:197, **206–208**
 advantages, **1:208**
 examples of, **1:206**
 history of, **1:206**
 key features of, **1:206–207**
 limitations of, **1:208**
 problem-solving with, **1:207–208**
Curse of dimensionality, **1:209–211**, 264
Curvilinear regression, **1:211–215**
 assumptions of, **1:214**
 confidence interval, **1:214**
 nonlinear regression, **1:214–215**
 plotting data, **1:212**
 polynomial regression, **1:212–213**
 power function evaluation, **1:213–214**
 uses, **1:211–212**
Cutoff score, **1:255**
Cut score, **1:9**, **2:495**
Cycle, **3:1005**

Dagnelie, P., **2:540**
D'Agostino, Ralph, Jr., **3:804**
D'Agostino, Ralph, Sr., **2:501**, **3:804**
Dale, A., **3:874**
Dalton, L. E., **1:116**
Daly, J., **3:837**
Damage modeling, **1:99**, **2:774**
DAMBE (Data Analysis in Molecular Biology and
 Evolution), **1:93**
Daniel, M. H., **1:50**

Darwin, Charles, **1:217–218**, 383, 384
Darwin, Erasmus, **1:383**
Das Gupta, Somesh, **1:95**
Data
 categorical, **1:43**, 70
 censored, **1:125–126**, **3:982–983**
 coding, **1:419**
 definition, **1:219**
 exploratory data analysis, **1:328–331**
 geostatistical, **3:925–926**
 incomplete, **1:310**
 lattice, **3:926**
 necessity of, **1:220**
 outdated, **1:319–320**
 point pattern, **3:924–925**
 primary versus secondary, **1:220**
 quantitative versus qualitative, **1:220**
 smoothing, **2:633–637**
 spatial, **3:924–926**
 tangible versus intangible, **1:220**
 training, **1:267**, 268
 truncated, **3:982**, 983
 types, **1:220**
 See also **Missing data method**
Data analysis
 grounded theory, **1:419**
 measurement versus statistics, **2:583**
 multiple comparisons, **2:644–647**
Data Analysis ToolPak, **1:218–219**
Data Archive, **3:873**
Databases
 commercial, **3:873**
 education, **3:873–874**
 government, **3:873**
 hard copy, **3:873**
 nongovernment, **3:872–873**
 poll service, **3:872–873**
 secondary data analysis and, **3:872–874**
Data collection, **1:219–225**
 Delphi technique, **1:240**
 descriptive research, **1:252**
 ecological momentary assessment, **1:298**
 effective, strategies for, **1:223–225**
 ethical considerations, **1:221–222**
 impediments, **1:222–223**
 methods, **1:220–221**
 necessity of data, **1:220**
 secondary data analysis, **3:874**
 significance of, **1:219**
 structural equation modeling, **3:974**

- “Ten Commandments” of, 3:1059–1060
types of data, 1:220
- Data compression, 1:225–231**
application areas, 1:226
compression approaches, 1:226
data modeling approaches, 1:226–228
definition, 1:225–226
lossless versus lossy, 1:226
source-of-information modeling, 1:228
users-of-information modeling, 1:228–231
- Datamax Corporation, 1:188
- Data mining, 1:232–235**
amount of data, 1:232
clustering, 1:235
computational statistics, 1:170
decision trees, 1:233–234
definition, 1:232
exploratory versus practical modeling, 1:232–233
genetic algorithms, 1:234
goals, 1:232–233
modeling procedures, 1:233–235
nearest neighbor methods, 1:234–235
neural networks, 1:234
origins of, 1:233
process, 1:233
rule induction, 1:234
statistical methods, 1:235
support vector machines, 1:234
terminology, 1:233
text mining, 1:235
traditional statistical analysis versus, 1:232–233
types of data, 1:232
uses, 1:232
- Data modeling, 1:226–228
- Data reduction, 1:210–211, 2:788
See also **Dimension reduction**
- Data structure, 1:225–226
- David and Carol Myers Foundation, 1:49
- Davis, S., 1:20
- Dayton, C. Mitchell, 2:501
- Dear, Robert E., 2:608–609
- De Castro, J. M., 1:212–213
- Decision boundary, 1:235–236**
- Decision making
Career Development Inventory, 1:119
Career Maturity Inventory, 1:120
conditions for, 1:236–237
Delphi technique, 1:240–245
ethical, 1:325
hypothesis testing, 2:697–698
nondirectional tests, 2:713
secondary data analysis, 3:875
- Decision Making Inventory (DMI), 1:121
- Decision-theoretic school of hypothesis testing, 2:697–698
- Decision theory, 1:236–240**
discriminant analysis, 1:268
example of, 1:238–240
game theory versus, 1:237
origins of, 1:237
statistical, 1:237–238
- Decision time/speed, 1:3
- Decision trees, 1:233–234
- Declaration of Helsinki (1964), 2:466
- Defining relation, 1:376
- De la Rosa, I. A., 1:116
- Delphi technique, 1:240–246**
advantages, 1:243
anonymity, 1:241
consensus theory, 1:241
disadvantages/limitations, 1:243–244
divergent thought, 1:241
ethical considerations, 1:243–244
historic, 1:242, 244–245
mixed methods, 1:242
Nominal Group Technique and, 1:242, 245
policy, 1:242
process, 1:241–242
Q-Sort technique and, 1:245
reactive, 1:243
real-time, 1:242
research applications, 1:244
research purposes and uses, 1:241
statistical measures of agreement, 1:241–242
theoretical basis, 1:241
variations, 1:242–243
- Delta method, 1:246–248**
- Dembo, Amir, 3:865
- Demetriou, G., 1:94
- Deming, William Edwards, 1:248–249**
- Dempster, A., 1:310
- Deno, Stanley, 1:206
- Denograms, 2:503
- Denotative meaning, 3:878
- Density estimation, 3:919
- Deontological tradition, ethics in, 1:323
- Dependence measures, 1:102
- Dependent samples *t* test. *See* **Paired samples *t* test**
- Dependent variable, 1:249–250**
characteristics, 1:249, 3:1036

- experimental research, 2:624
- independent and, 3:1037
- multivariate analysis of variance, 2:669
- research examples, 1:249–250
- types of measurement represented by, 1:249
- Depression**
 - Beck Depression Inventory, 1:82–83
 - Carroll Depression Scale, 1:121
 - Clinical Assessment of Depression, 1:148–149
- Descartes, René, 2:745
- Descriptive research, 1:250–253**
 - diaries, 1:253
 - focus groups, 1:252
 - in-depth interviews, 1:252
 - methods, 1:251–253
 - observation, 1:251–252
 - personal journals, 1:253
 - qualitative research and, 1:250–251
 - quantitative research and, 1:250–251
 - questionnaires, 1:251
 - structured interviews, 1:251
 - utility of, 1:253
- Design data. *See* Training data
- Detail balance, 2:570
- Determinant, of matrix, 1:307–308
- Deterministic causality, 1:124
- Detroit Tests of Learning Aptitude, 3:846
- Detterman, D. K., 1:50
- Deviation, 1:72, 328
 - See also* **Standard deviation**
- Deviation IQ, 2:475
- Deviation score, 1:253–254**
 - See also* **Standard deviation**
- Dewey, Davis R., 2:500
- d* family, 2:596
- Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association), 1:82, 121, 145, 147, 148, 255, 319, 2:512, 606, 766, 792, 3:976–977
- Diagnostic validity, 1:254–258**
 - See also* **Validity theory**
- Diagonalization, 1:307
- Diagrams. *See* **Graphical statistical methods**
- Diaries, 1:253, 3:878
- Dichotomy, 1:87
- Die tosses, 2:627
- Dietz, E. Jacquelin, 2:502
- Di Fazio, R., 2:430
- Difference-based scaling, 3:864
- Difference engine, 1:69
- Difference score, 1:258–260**
 - advantages, 1:259
 - ANCOVA versus, 1:259
 - appropriateness for mean changes, 1:259
 - appropriateness for predictors of change, 1:259
 - definition, 1:258
 - disadvantages, 1:259
- Differential Aptitude Test (DAT), 1:260–261**
 - development of, 1:2
 - overview of, 1:4
- Differential bundle functioning (DBF), 2:734
- Differential equations, 1:86
- Differential item functioning (DIF), 2:490–491, 734
- Differential Personality Inventory, 1:71–72, 2:442
- Differential prediction, 2:492
- Differential test functioning (DTF), 2:734
- Diffusion limited aggregation (DLA), 1:373
- Diggle-Kenward Model for Dropout, 1:261–264**
- Dignity, in research, 1:324–325
- Dimensional scaling, 1:395
- Dimension reduction, 1:264–267**
 - benefits, 1:264
 - case study, 1:266–267
 - methodologies, 1:265–266
 - supervised, 1:265–266
 - unsupervised, 1:265
- Direct attitude measures, 1:53–54
- Direct correlation, 3:938
- Directed divergence, 1:15
- Directional alternative hypothesis, 1:23–24
- Directional hypothesis, 2:459
- Dirichlet, Gustav Peter Lejeune, 2:772
- Disability
 - active life expectancy, 1:11
 - alternate assessment, 1:19–20
 - definition, 1:27
 - See also* **Americans with Disabilities Act**; Individuals with Disabilities Education Act; Individuals with Disabilities in Education Improvement Act; **Section 504 of the Rehabilitation Act of 1973**
- Disability-free life expectancy (DFLE), 1:11
- Discordant bivariate, 1:102
- Discourse assessment, 2:782–783
- Discrete Cosine Transform (DCT), 1:229
- Discrete-event simulation models, 3:901–903
- Discrete Fourier transform, 1:365, 367–369
- Discrete models, 3:901
- Discrete variable, 1:184–185
- Discriminability scaling, 3:862–863
- Discriminant analysis, 1:267–270**
 - allocation rules, 1:267–268

- decision boundaries, 1:236
- decision theory and, 1:268
- discriminant correspondence analysis, 1:270–275
- flexible rules, 1:269–270
- sample-based allocation rules, 1:268–269
- Discriminant correspondence analysis, 1:270–275**
- Discriminant function coefficients, 2:671
- Discriminant validity, 2:666
- Disjunctive model, 1:163
- Dispersion, 1:66
- Dispersion test statistic, 2:773
- Dissertation Abstracts International (DAI)*, 1:25
- Dissertations, 1:24–25
- Dissimilarities, 2:604
- Dissimilarity coefficient, 1:275–280**
- Distance, 1:280–284**
 - Euclidean, 1:281
 - generalized Euclidean, 1:281–283
 - Hellinger, 1:283–284
 - matrix analysis, 1:284
 - Minkowski's, 1:283
 - profile analysis, 2:788
 - significance of, 1:280
 - sorting, 1:280
- Distance matrices
 - DISTATIS, 1:284
 - metric multidimensional scaling, 2:598–604
 - transforming, to cross-product matrices, 2:599–601
- DISTATIS, 1:284–290**
 - compromise matrix, analysis of, 1:288–289
 - compromise matrix, computation of, 1:286–288
 - example of, 1:285–290
 - goal of, 1:284
 - process, 1:285
 - projection into compromise space, 1:289–290
 - R_v and congruence coefficients, 3:849–852
- Distractors, in multiple-choice tests, 2:643–644
- Distributive justice, in research, 1:324
- Divergence hypothesis, 1:50
- Divergent thinking, 1:241, 2:477, 3:1012
- Diversity coefficient, 1:278–280
- Dixon test for outliers, 1:291–293**
- DNA. *See* Genetic data
- Dobson, Annette J., 1:16
- Doctoral dissertations, 1:24–25
- Doctoral Dissertations Accepted by American Universities*, 1:25
- Doll, Edgar A., 3:1045
- Dol* pain scale, 3:862
- Domain definition, 1:181
- Domain-referenced interpretations. *See* Criterion-referenced interpretations
- Domain-referenced tests, 1:197
- Domain relevance, 1:182
- Domain representation, 1:181–182
- Dominican order, 2:707–708
- Dominick, J. R., 3:874
- Doten, Carroll W., 2:500
- Double-barreled questions, 3:809
- Double principal coordinate analysis, 1:278
- Dougherty, Donald M., 2:452, 525
- Draw a Man Test, 1:406
- Draw a Person Test, 2:791
- Draw a Woman Test, 1:406
- Dr Foster (organization), 3:831–832
- Dropout. *See* Attrition bias; Diggle-Kenward Model for Dropout
- Drug-Free Workplace Act (1988), 1:27–28
- Du, Y., 2:688
- Dual discrepancy, 3:841–842
- Dual scaling, 2:651
- Dubanoski, J., 3:1005
- Dunbar, S. B., 2:759
- Dunham, R. B., 1:35
- Dunn, Olive Jean, 1:293, 294
- Dunn approximation, 1:105
- Dunnett's test, 2:647
- Dunn-Šidák multiple comparison test, 1:294
- Dunn's multiple comparison test, 1:293–295, 2:646**
- Durham, T. W., 1:83
- Durlak, Joseph A., 2:597
- Dynamic systems, 3:901
- Ear, data compression for, 1:229
- Eberstadt, Nicholas, 1:69
- Ecological momentary assessment (EMA), 1:297–298, 3:878**
- Edelman, B., 1:284
- EdITS/Educational and Industrial Testing, 2:641
- Education, authenticity and, 1: 61–62
- Educational and Industrial Testing Service, 1:175
- Educational Measurement: Issues and Practice*, 2:678
- Educational Testing Service (ETS), 1:173, 298–299, 410**
- Education databases, 3:873–874
- Education for All Handicapped Children Act, 2:454–456
- Edwards, Allen, 1:299
- Edwards, J. R., 1:260
- Edwards Personal Preference Schedule (EPPS), 1:299–300**
- Effect size, 1:300–304**

- correlation and proportion of variance measures, 1:301–302
- d* family, 2:596
- estimation of, 2:596–597
- indices, properties of, 1:300
- indices, types of, 1:301–302
- inferential statistics, 2:460
- interpretation, 1:302–303
- programs for calculating, 1:303
- research process role of, 1:300
- r* family, 2:596
- sample size, 3:859–860
- selected test statistics, 1:302
- significance of, 1:300–301
- standardized mean difference, 1:301
- Effect Size Calculator, 1:303
- Effect Size Determination Program, 1:303
- Eigendecomposition, 1:304–309**
 - definition, 1:304
 - determinant, 1:307–308
 - notations and definition, 1:304–306
 - positive semi-definite matrices, 1:306–307
 - rank, 1:308
 - significance of, 1:304
 - singular value decomposition, 3:907–912
 - statistical properties, 1:308–309
 - trace, 1:307
- Eigenvalue criterion method, 3:868
- Eigenvalues, 1:309–310**
 - eigendecomposition, 1:304–309
 - exploratory factor analysis, 1:333–334
 - multiple correspondence analysis, 2:653–654
- Eigenvalues-greater-than-1 rule, 1:333
- Eigenvectors, 1:304–306, 309–310
- Eight-Year Study, 1:203
- Elderly Memory Schedule, 3:846
- Elementary and Secondary Education Act, 1:19
- Elmore, P. B., 3:850–851
- Elsholtz, Johann Sigismund, 1:35–36
- EM algorithm, 1:310–313**
 - discriminant analysis, 1:269
 - expectation in, 1:310
 - maximization in, 1:310
 - survival analysis, 3:984
- Embedded assessment, 2:776
- Embedded Figures Test (EFT), 1:313–314**
- Embretson, Susan, 3:1035
- Emerson, Ralph Waldo, 3:1029
- Emotion, 1:217
- Emotionally Focused Therapy, 1:73
- Empiricism, Ockham's razor and, 2:708
- Empty model, 2:434
- Endogenous explanatory variables, 2:468–469, 746, 3:974
- English language learners, 1:20
- Enlisted Testing Program (ETP), 1:45–46
- Epidemiology, 1:91–92, 412, 3:833
- Equal-appearing intervals method, 1:53, 2:536–537
- Equal correlation, 1:401
- Equality of means testing, 1:84–85
- Equivalence testing, 1:314–315**
- Erdős, Paul, 1:383
- Ergodic theorem, 2:569–570
- Ergonomics, 1:37
- Error, 3:944–945
 - random, 1:141, 2:585, 730, 3:933
 - sampling, 3:861–862
 - true score and, 3:1014–1016

See also **Measurement error; Standard error of measurement; Standard error of the mean**
- Error of the second kind, 3:1021
- Error prediction, 2:546
- Error ratio, 1:423
- Escoufier, Y., 3:849, 955
- Essay items, 1:315–316**
 - advantages, 1:315–316
 - disadvantages, 1:316
 - guidelines for writing, 1:315
 - open- versus closed-ended, 1:315
 - scoring, 1:316
- Essay tests, 2:460
- Essentially tau-equivalent tests, 1:142
- Essex, C., 1:128, 2:585
- Estimate of reliability, 1:155
- Estimates of the population median, 1:316–318**
- Estimation
 - area of irregular figure, 2:628
 - area of regular figure, 2:627–628
 - definition, 2:457
 - eyeball, 1:335–336
 - generalized estimating equations, 1:397–403
 - hierarchical linear modeling, 2:434–435
 - inferential statistics, 2:457–459
 - instrumental variable, 2:470–472
 - irrational numbers, 2:628–629
 - Kaplan-Meier, 1:126, 3:983–985
 - magnitude, 3:863–864
 - maximum likelihood method, 2:575–576
 - method of moments, 1:390–394
 - missing data method, 2:608–612
 - mixed-effects models, 2:614–615

- Monte Carlo methods, 2:627–629
 random sampling and, 3:817–818
 statistical decision theory and, 1:237, 238
 structural equation modeling, 3:974–975
 unbiased estimator, 3:1025–1026
- Estimation of the unknown b , 1:168
- Estimator variables
 criterion validity and, 1:200, 202
 decision theory and, 1:238–240
 ratio, 1:248
- Ethical issues in testing, 1:318–321**
 evaluator qualifications, 1:320
 informed consent, 1:319
 scoring and interpretation, 1:320
 selection and uses of tests, 1:318–319
 test design, 1:319–320
 test security, 1:320–321
 use of results, 1:320
- Ethical principles in the conduct of research with human participants, 1:321–325**
 basic, 1:323–325
 cost-benefit considerations, 1:323
 decision making, 1:325
 history of, 1:322
 informed consent, 2:465–468
 internal review boards, 2:480–483
 philosophical approaches, 1:322–323
 responsibilities in research, 1:323
- Ethics
 bias and, 2:489
 confidentiality, 1:319
 data collection and, 1:221–222
 definition, 1:321
 Delphi technique and, 1:243–244
 informed consent, 1:319
 projective testing, 2:792
 secondary data analysis, 3:874
See also **Ethical issues in testing; Ethical principles in the conduct of research with human participants**
- Ethnicity
 ability tests and, 1:4–5
 System of Multicultural Pluralistic Assessment, 3:988
See also Race
- Euclidean distance, 1:281, 2:599
- Eugenics, 1:384
- Evaluation dimension of connotative meaning, 3:878, 881
- Evaluation of special functions, 1:169
- Evaluative priming technique, 1:54–55
- Evaluators, qualifications of test, 1:320
See also Interrater reliability
- Evans, M. G., 1:48
- Event-contingent schedules, 1:297
- Event related brain potentials (ERP), 1:55
- Evidence-based practice (EBP), 1:326**
- Evolution, 1:217
- Exact identification, instrumental variables and, 2:470
- Exact matching, 2:527, 3:827–828
- Exact probability, 1:89, 137
 Fisher exact probability test, 1:354–358
- Excel Data Analysis ToolPak, 3:1043
- Excel spreadsheet functions, 1:326–328, 361, 3:935–937**
- Excel VAR function, 3:1042
- Exclusive range, 3:820
- Exit exams, 2:436
- Exner, John E., 3:848
- Exogenous explanatory variables, 2:468–469, 746, 3:974
- Exon, 1:91
- Expectation-conditional maximization, 1:313
- Expectation-conditional maximization either, 1:313
- Expectation-Maximization algorithm. *See* **EM algorithm**
- Expectation-Minimization algorithm, 2:617
- Expected a posteriori estimate, 1:80, 2:781
- Experience sampling method, 1:297
- Experiments
 ANCOVA and, 1:29
 causality and, 1:124
 observational studies versus, 2:704–705
- Experimentwise alpha, 1:104
- Expert networks, 2:621
- Explained variance, 3:910
- Explanation, regression and, 3:831
- Exploratory data analysis (EDA), 1:328–331**
 definition, 1:328
 development of, 1:329
 example of, 1:329–331
 graphics for, 1:413
 history of, 1:328–329
 scree plots, 3:868–869
- Exploratory Data Analysis* (Tukey), 1:328–329
- Exploratory factor analysis (EFA), 1:332–335, 342–343**
 confirmatory versus, 1:332, 344
 decisions in conducting, 1:333–334
 example of, 1:332–333
 mathematical basis, 1:332
 model fitting, 1:334, 343
 number of common factors, 1:333–334, 342–343
 rotating a solution, 1:335, 343
- Exploratory modeling, 1:232–233

- Exploratory spatial data analysis (ESDA), 1:329
Exploring Careers, 1:46
Extended Reading and Math Tasks, 1:21
External validity, 3:806
Exxon Education Foundation, 2:502
Eye, data compression for, 1:229–230
Eyeball estimation, 1:335–336
Eysenck, Hans J., 2:595
- Face-to-face interviews, 3:809
Face validity, 1:337–340
 advantages, 1:337–338
 appearance and, 1:339
 assumption and, 1:339
 content validity and, 3:1034
 definition of, 1:337
 disadvantages, 1:338
 expert determination of, 1:339
 history of, 1:338–339
 hypothesis and, 1:339
 meta-analyses and, 1:339–340
 performance-based assessment, 2:759
 significance of, 1:337
Factor analysis, 1:341–345
 ability tests and, 1:2
 causal analysis and, 1:125
 common factor model, 1:341–342
 Comrey Personality Scales, 1:175
 confirmatory, 1:343–344
 construct validity, 1:178
 data reduction, 1:211
 definition, 1:341
 dimension reduction, 1:265
 exploratory, 1:332–335, 342–343
 multiple, 2:657–663
 purpose, 1:345
 Q methodology, 3:801–802
 scree plots, 3:868–869
Factored homogenous item dimension, 1:174, 175
Factor extraction, 1:334, 343
Factorial ANOVA, 1:33, 34
Factorial design, 1:348–351, 3:892
Factorial MANOVA, 2:669
Factorial repeated measures design, 2:559
Factor invariance, 2:491–492
Factor loading, 2:784
Factor score coefficients, 1:345–346
Factor score indeterminacy, 1:346–348
Factor scores, 1:345–348
 eigendecomposition, 1:308
 evaluating, 1:347–348
 explanation of, 1:345–346
 refined and coarse, 1:346–347
Factor solution, 1:178
Fagan, Joseph F., III, 1:351
Fagan Test of Infant Intelligence (FTII), 1:351–352
Fahoom, G., 2:630
Fail-Safe-N, 1:353, 2:597
Failure time data, 3:982–987
FairTest, 1:299
Faking Bad Bias, 2:430
Faking Good Bias, 2:430
Fallible score, 1:130
Fallin, K., 1:20
False negatives, 1:255, 3:1021
False positives, 1:255, 2:697–698
Family Education Rights and Privacy Act, 2:455
Family Environment Scale (FES), 1:352–353
Family Kinetic Drawing Test, 2:791
Family of tests, 1:104
Familywise alpha, 1:104, 2:644–646
Familywise error rate. *See* Familywise alpha
Fantuzzo, J. W., 2:454
Farver, J. A. M., 3:1050
FASTA (fast algorithms), 1:93
Fazio, Russell, 1:54
Fechner, Gustav Theodor, 3:862
Federal Aviation Administration, 1:114, 2:463
Federal Policy for the Protection of
 Human Subjects, 2:481
Fedworld, 3:873
Feedback loops, 3:974
Feed-forward net, 1:46–47
Feldt, Leonard S., 1:157
Fermat, Pierre de, 1:129, 2:745
Fermi, Enrico, 2:627
Ferrari, J. R., 1:56
Ferrer-Caja, E., 2:475
Ferris, K. R., 2:554
Ferron, J. M., 1:344
Field, A. P., 2: 523
Field, S., 1:179
Fienberg, S. E., 3:894
Figure drawing projective tests, 2:791
File drawer problem, 1:353, 2:596
Fill's Algorithm, 2:573
Finch, S., 1:26
Finite mixture distributions, 2:616–617
Firing rule, 1:46–47
A First Course in Factor Analysis (Comrey), 1:174, 175

- First singular value, 2:657
- Fisher, John Dix, 1:27
- Fisher, Ronald Aylmer**, 1:32, 268, **354**, 358, 359, 360, 2:457, 530, 706, 750, 773, 3:942, 1051
- Fisher exact probability test**, 1:137, **1:354–358**
- Fisher information, 1:172
- Fisher-Irwin test. *See* **Fisher exact probability test**
- Fisher's LSD**, 1:359–360, 2:646–647, 681
- Fisher's Z transformation**, 1:360–364
- averaging correlations, 1:363–364
 - confidence intervals for correlations, 1:362
 - correlations from independent samples, 1:363
 - hypothesis testing, 1:361–362
- Fisher transform, 1:190
- Fiske, D., 1:179, 2:666
- Fitted candidate models, 1:15
- Fitzpatrick, J., 3:874
- Fixed-effect models, 1:272–275, 399
- Fleiss, Joseph, 2:579
- Fletcher, Richard, 1:27
- Flexible discriminant rules, 1:269–270
- Fluid intelligence (Gf)
- Culture Fair Intelligence Test, 1:204–205
 - definition, 1:2, 204, 405
 - Stanford-Binet Intelligence Scales, 3:951
 - theory of, 1:404, 2:477
- Focal groups, 2:490
- Focus groups, 1:252, 3:813
- Forced distributions, 3:800
- Forced-free distributions, 3:800
- Ford, Harrison, 2:497
- Forecasting
- Delphi technique, 1:240–245
 - moving average, 2:633–638
 - See also* Prediction
- Forensic science, 1:37
- Form Alpha, 2:477
- Formative achievement tests
- curriculum-based measurement, 1:207
 - summative versus, 1:9
- Form Beta, 2:477
- Formula 20, 1:155
- Formula 21, 1:155
- FORTTRAN, 2:627
- Forward stepwise regression, 1:132
- Fourier, Jean Baptiste, 1:365
- Fourier analysis, 1:365, 374–375
- Fourier transform**, 1:364–369
- continuous, 1:366
 - definition, 1:364
 - discrete, 1:367–369
 - frequency, 1:365
 - history of, 1:365
 - types, 1:365
- Fourier transform pair, 1:365
- Fox, C., 2:583–584
- Fractal**, 1:370–375
- definition, 1:370
 - diffusion limited aggregation, 1:373
 - fractal dimension, 1:374
 - fractal functions, 1:371–372
 - iterated function systems, 1:370–372
 - random processes, 1:373–374
 - self-similarity, 1:370–373
 - strange attractors, 1:373
 - wavelets, 1:374–375
- Fractional Bayes factors, 1:75
- Fractional randomized block design**, 1:375–379
- advantages, 1:379
 - computational example, 1:379
 - constructing, 1:376–379
- Francia, R. S., 3:884
- Franciscan order, 2:707–708
- Frankfurt Convention (1882), 1:36
- Frankfurt horizontal plane, 1:36
- Frazer, G. H., 1:244
- French Academy of Sciences, 2:709
- Frequency, 1:365
- Frequency distribution**, 1:380
- class interval, 1:139–140
 - cumulative, 1:205–206
 - grouped, 2:586
 - histograms, 2:440
 - kurtosis, 2: 522–523
- Frequentist analysis, 1:80–81, 3:942
- Freud, Sigmund, 2:768
- Freund, Y., 1:270
- Friedman, Milton, 1:380, 2:722
- Friedman test**, 1:151, **380–382**, 2:722
- Friedman two-way analysis of variance, 2:687
- Frobenius scalar product, 3:850
- F* test
- analysis of variance, 1:32–34, 3:836–837
 - critical values table for, 3:1066–1069
 - one-way analysis of variance, 2:713–716
 - partial regression coefficient, 2:738
 - Peritz procedure, 2:760–762
 - sample size, 3:858–859
 - software packages, 3:837
 - Tukey-Kramer procedure, 3:1016

- Fuchs, L., 1:208, 3:841
 Fullerton Longitudinal Study, 1:139
 Full information maximum likelihood, 2:471
 Full rank matrix, 1:308
 Full-scale IQ, 2:477
 Functional magnetic resonance imaging (fMRI), 1:55–56
 Functional space, 1:265
 Fundamental, 1:367
 Fund for the Improvement of Postsecondary Education, U.S. Department of Education, 2:502
 Funnell plots, 2:597
 Future Problem Solving Program (FFSP), 3:1010
 Fuzzy central limit theorem, 1:127
- g. See* General intelligence (*g*)
 Gain score, 1:258
 Gaizauskas, R., 1:94
 Galileo Galilei, 2:530, 690
 Gallaudet Research Institute, 3:950
 Gallup Poll, 3:872–873
Galton, Sir Francis, 1:37, 1:217, 383–384, 2:441, 473, 478, 530, 710, 750
Gambler's fallacy, 1:384–385
 Games, Paul, 1:294
 Game theory, 1:237
 Gandhi, Mohandas, 1:244
 Garb, H. N., 3:848
 Gardner, Howard, 1:403
 Garn, Stanley, 3:799
 Gating networks, 2:621
 Gauge repeatability studies, 3:916
Gauss, Carl Friedrich, 1:385–386, 2:530, 690, 691, 701, 3:1025
 Gaussian distribution, 2:619–620, 622
See also Normal distribution
 Gauss-Markov conditions and theorem, 2:531
 Gauss-Newton approximations, 2:532
 Gc. *See* Crystallized intelligence
 Gee, Travis, 2:765
 Gender
 ability tests and, 1:4–5
 Male Role Norms Inventory, 2:563
 Gene expression, 1:90
 General Aptitude Test Battery (GATB)
 controversy over, 1:2, 4
 occupational use of, 1:41
 General Clerical Test, 2:608
 General health policy model, 3:804
 General intelligence (*g*)
 ability tests and, 1:1
 asymmetry of, 1:50
 Cattell-Horn Gf-Gc theory of intelligence, 1:404, 3:1053–1054
 Gf-Gc theory of intelligence, 1:403–406
 Raven's Progressive Matrices, 3:826–827
 reaction time and, 2:478
 Spearman's theory of, 2:477
 Stanford-Binet Intelligence Scales, 3:951
 Universal Nonverbal Intelligence Test, 3:1026–1027
 Generalizability theory (G Theory)
 classical test theory and, 1:142
 Cronbach/Gleser rewriting of, 1:204
 error in, 3:1016
Generalized additive model, 1:386–389
Generalized estimating equations, 1:397–403
 example of, 1:401–402
 generalized linear models, 1:398
 independence model, 1:398–399
 software packages, 1:400
 subject-specific versus population-averaged models, 1:399–400
 working correlation matrix estimation, 1:400–401
 Generalized Euclidean distance, 1:281–283
 Generalized expectancy scale, 2:479
 Generalized instrumental variable (GIV)
 estimation, 2:470
 Generalized interaction, 1:377
 Generalized least squares, 2:531
 Generalized linear mixed models, 2:615
 Generalized linear models, 1:386, 398, 2:620
Generalized method of moments, 1:390–394
 application of, 1:392–394
 examples of, 1:392
 identification and overidentification, 1:390
 overidentification test, 1:392
 weighted and optimal, 1:390–391
 Generalized method of moments (GMM) estimator, 1:390–394
Generalized Procrustes Analysis (GPA), 1:394–397
 Generalized treatment, 1:377
 General linear models, 2:544–545
See also **Generalized additive model**
 General linear statistical model, 1:194
 General-purpose programming language, 3:903
 General Social Survey, 3:872
 Generative causation, 3:938–939
 Genetic algorithms, 1:234
 Genetic data, 1:90–94, 3:866–867
 Genuine correlation, 3:938

- Geoffroy, J., 1:98
 Geographical Information Systems, 3:926
 Geometric mean, 1: 65, 2:587–588, 591
 Geostatistical data, 3:925–926
 Gerberding, Julie Louise, 1:126
Gerontological Apperception Test, 1:403
 Gershater-Molko, R. M., 3:819
 Gertrude Cox Scholarship, 1:27
 Gf. *See* Fluid intelligence
Gf-Gc theory of intelligence, 1:403–406
 abilities, 1:404–406
 development of, 1:404
 Gibbs sampler, 2:572
 Gibson, C. L., 3:1043
 Glaser, Barney, 1:418, 420
 Glaser, R., 1:195
 Glass, Eugene (Gene), 1:97, 2:595
 Glass's Δ , 2:596
 Glaz, J., 3:865
 Gleason, Terry, 2:608
 Gleser, Goldine, 1:157, 204
 Global data matrix, 2:658–659
 Goddard, Henry, 2:476
 Gold standard, 1:255, 257–258
 Good, Philip I., 2:501
Goodenough Harris Drawing Test, 1:406
 Goodness of fit
 Akaike information criterion, 1:15
 exploratory factor analysis, 1:334
 multivariate normal distribution, 2:674
 Q-Q plot, 3:802–804
 secondary data analysis, 3:874
 structural equation modeling, 3:975
 Goodness of fit index, 3:975
Goodness-of-fit tests, 1:407–410
 illustration of, 1:409–410
 Kolmogorov-Smirnov test, 1:408
 purpose, 1:407
 Shapiro-Wilk test for normality, 3:884
 structural equation modeling, 3:975
 See also **Chi-square test for goodness of fit**
 Gorin, Joanna, 3:1035
 Gosset, William Sealy, 2:627, 3:990
 Gough, Harrison, 1:12, 117
 Gould, Stephen Jay, 1:218
 Government databases, 3:873
 Grade-to-grade promotion exams, 2:436
 Gradient descent, 2:532
 Graduate Management Admission Test (GMAT), 1:171, 299
 Graduate Record Examinations Board, 1:411
Graduate Record Examinations (GRE), 1:171, 173, 298, 410–411
Grand mean, 1:411–412
 Grand mean centering, 2:435
 Grand variance, 2:714
Graphical statistical methods, 1:412–416
 birth of modern, 1:413–414
 founders of, 1:412–413
 interactive, 1:414–416
 Graves, E. J., 3:869, 874
 GRE. *See* **Graduate Record Examinations**
 Greenacre, M. J., 2:653
 Greene, William, 2:471
 Greenhouse-Geisser estimate, 1:35
 Greenwood formula, 3:984
 Gregory, James, 3:896
Gresham, Frank M., 1:417
 Grissom, R. J., 1:301
Grounded theory, 1:418–421
 criticisms of, 1:419–420
 history of, 1:418
 philosophical basis, 1:418–419
 procedures, 1:419
 purposes, 1:418
 Group differences
 ability tests, 1:4–5
 aptitude tests, 1:42
 item and test bias, 2:489–490
 Universal Nonverbal Intelligence Test, 3:1026–1027
 Grouped frequency distribution, 2:586
 Grouping variable, 3:990
 Group of pictures (GOP), 1:230–231
 Groups mean centering, 2:435
 Group variable, 3:990
 Growth modeling. *See* **Longitudinal/repeated measures data**
 Guessing parameter, 2:493
 Guest, G., 1:424
 Guilford, J. P., 1:142, 174, 2:477, 3:1010
 Guinness, Ruth Eileen, 1:354
 Gulliksen, H., 1:142
 Gutin, B., 1:38
 Guttman, Louis, 1:53, 421
Guttman scaling, 1:53–54, 421–424, 3:880
 criteria, 1:421–422
 evaluating reproducibility, 1:422–423
 purpose, 1:421
 strengths/weaknesses, 1:423

- Hahn, C.-S., 1:168
Haig, Brian, 1:419
Hakstian, A. Ralph, 1:157
Haladyna, Thomas, 2:642, 643
Hall, D. G., 1:166
Hamada, W. C., 3:1005
Hamagami, F., 2:475
Hamilton Depression Rating Scale (HAMD), 1:121, 2:769
Hamming distance, 1:283
Hammond, Kenneth, 1:54
Hancock, Gregory R., 2:501
Handbook of Test Development, 1:8
Hand Test, 2:789
Hanes, J., Jr., 1:38
Hansen, Lars Peter, 1:390
Hans Huber, 3:848
Hanson, David Lee, 2:501
Harcourt Assessment, 1:81, 2:573, 3:949, 950, 1000, 1048
Hardin, E. E., 1:121
Hardlimit firing rule, 1:47
Harmonic mean, 1: 65, 2:425–426, 588
Harrell-Davis estimator, 1:317–318, 2:593
Harris, H., 3:875
Harris, Richard J., 2:670
Harrison, T. C., 1:28–29
Harter Teacher Rating Scale, 3:921
Hartigan, J. A., 2:631
Harvard University Press, 3:1000
Hastie, T., 2:785
Hastings, W. K., 2:571
Hathaway, Starke, 2:609
Haughton, B., 1:244
Hawthorne Effect, 2:430
Haxby, J. V., 2:598
Haynes, O. M., 1:168
Hayter, A. J., 1:359–360, 3:1016
Hayter-Fisher version of Tukey-Kramer, 3:1017–1018
Hazard function, 3:982
Headrick, Todd C., 2:501
Health-adjusted life expectancy (HALE), 1:11
Healthcare Common Procedure Coding System, 2:429
Health History Inventories, 3:988
Health informatics, 1:90
Health Insurance Portability and Accountability Act
 (HIPAA), 2:426–429
 administrative simplification, 2:427
 Identifier Rule, 2:429
 portability, 2:426–427
 Privacy Rule, 2:427–428
 purpose, 2:426
 Security Rule, 2:428–429
 Transaction and Code Sets Rule, 2:429
Healy, M. J. R., 2:674
Heaton, Janet, 3:872, 874
Heavy-tailed distribution, 2: 522
Heckman, J. J., 1:59–60
Hedges, Larry V., 2:595, 597
Hedges's and Olkin's *d*, 2:597
Hedges's *g*, 2:596
Hellinger distance, 1:283–284, 2:656
Hello-Goodbye Effect, 2:429–431
Hempel, Carl G., 3:1034
Henry, S., 1:125
Henson, R. K., 1:15
Henze, N., 2:674
Henze Zirkler test statistic, 2:674
Hereditary Genius (Galton), 2:441
L'Hermier des Plantes, H., 3:955
Herrnstein, Richard J., 1:45
Hertz, Heinrich, 3:1047
Heteroscedasticity and homoscedasticity,
 2:431–432, 546
Heterotrait heteromethod triangles, 2:667–668
Heterotrait monomethod triangles, 2:667–668
Heydenberk, R. A., 3:992
Heydenberk, W. R., 3:992
H. H. Wilson Company, 1:25
Hidden bias, 2:706
Hierarchical clustering, 1:235
Hierarchical linear modeling, 2:432–435
 applications of, 2:433
 assumptions of, 2:435
 centering Level-1 predictors, 2:435
 estimation, 2:434–435
 intraclass correlation coefficient, 2:435
 no-predictors two-level, 2:434
 purpose, 2:432
 repeated measures analysis versus, 2:561
 software packages, 2:435
 two-level, 2:433–434
Hierarchical multiple regression, 3:996–998
Higher Education Research Institute, 3:874
High school graduation exams, 2:436
High-stakes tests, 1:10, 2:435–439
 benefits, 2:437
 effects of, 2:436
 history of, 2:436–437
 negative effects, 2:437–438
 theory of, 2:436
Hilbert-Schmidt matrix scalar product, 3:850

- Hines, C. V., 1:344
 Hinges, 1:112
 Hirohito, Emperor, 1:248
Histogram, 2:439–441
 bar graph versus, 2:440
 cautions, 2:440–441
 definition, 2:439
 example of, 2:439–440
 exploratory data analysis, 1:330
 information in, 2:439
 types, 2:440
 Historic Delphi technique, 1:242, 244–245
Historiometrics, 2:441
 Hit, 2:697–698
 Hitier, S., 3:851
 Hit rate, 1:256
 Hoaglin, D. C., 2:774
 Hoffman, A., 1:179
 Hogarty, K. Y., 1:344
 Hoh, Josephine, 3:867
 Holden, Ronald R., 2:442
Holden Psychological Screening Inventory (HPSI), 2:442
 Holland, John, 3:971
 Hollander, Myles, 2:686
 Holland model, 1:118
 Holzinger, K. J., 2:546
 Homogeneity analysis, 2:651
Homogeneity of variance, 2:442–443
 O'Brien test, 2:701–704
 Homoscedasticity. *See Heteroscedasticity and homoscedasticity*
 Honestly significant difference (HSD) procedure, 1:359–360, 2:647, 681, 778–780, 3:1016
 Hooke, Robert, 1:86
 Hopfield net, 1:47
 Horn, John, 1:404, 2:477–478
 Horst, P., 3:980
 Horst, Paul, 2:609
 Hotelling, Harold, 2:783
 Hotelling's *T*, 2:669
 House-Tree-Person Test, 2:791
 Hoyt, Cyril J., 1:155
 Hrdlička, Aleš, 1:36, 37
 Huang, W. H., 1:95
 Hubbard, Kin, 2:563
 Huberty, C. J., 1:384
 Huffcutt, A. I., 1:335
 Huffman, David, 1:226
 Huffman coding, 1:226–227
 Human chromosomes, 1:93
 Human Genome Project, 1:91
 Humbles, P., 1:38
 Hume, David, 1:123
 Humphreys, K., 1:94
 Hunt, Earl, 2:478
 Hunter, I. M. L., 2:702
 Hunter, John E., 2:595
 Hurst exponent, 1:373
 Hurvich, Clifford M., 1:16
 Huyghens's theorem, 1:129
 Huynh-Feldt estimate, 1:35
 Hwang, F. K., 3:865
Hypergeometric distribution, 2:444–445
 approximations to, 2:445
 example of, 2:444–445
 Fisher exact probability test, 1:356–358
 formulas, 2:444
 Hyperplane, 3:978
Hypothesis and hypothesis testing, 2:445–448
 alternative hypothesis, 1:23–24
 basic research, 1:72
 equivalence testing, 1:314–315
 example of, 2:448
 Fisher's *Z* transformation, 1:361–362
 implications, 2:448
 inferential statistics, 2:457–458
 nomothetic approach, 2:684
 nonparametric methods, 2:685–688
 one- and two-tailed tests, 2:711–713
 outcome, 2:447–448
 outliers, 1:291
 problems in, 1:103
 procedure, 2:446, 458–459, 711, 3:1019–1020
 purpose, 2:445
 random generating process, 2:446–447
 sampling distribution, 3:860–861
 statistical decision theory and, 1:237, 238
 Type II error, 3:1019–1023
 IBM Corporation, 2:627
 IBM Intelligent Miner, 1:233
 Ice, G. H., 1:202
 Idiographic. *See Nomothetic versus idiographic*
Illinois Test of Psycholinguistic Abilities (ITPA-3), 2:451–452
 Illusory correlation, 3:938
 Image analysis, 1:348
 Image compression, 3:911–912
Immediate and Delayed Memory Tasks, 2:452–453

- Implicit Association Test (IAT), 1:55
- Implicit measures of attitude, 1:54–55
- Impulsivity measures, 2:525–526
- Imputation, 3:981
- Incidence rate restricted Poisson distribution, 2:773, 775
- Inclusive range, 3:820
- Incomplete data, 1:310
- Incomplete sentence projective tests, 2:791–792
- Incubation Model of Teaching, 3:1010
- Independence model, 1:399
- Independent ANOVA, 1:33
- Independent variable, 2:453–454**
 - characteristics, 2:454, 3:1036–1037
 - dependent and, 3:1037
 - experimental research, 2:624
 - multivariate analysis of variance, 2:669
- In-depth interviews, 1:252
- Index to American Doctoral Dissertations*, 1:25
- Indirect attitude measures, 1:54–55
- Indirect correlation, 3:938
- Individual change model. *See* **Longitudinal/repeated measures data**
- Individual differences
 - ability tests, 1:1
 - Darwinian theory, 1:217
 - Embedded Figures Test, 1:313–314
 - ethics and, 1:325
 - historiometrics, 2:441
 - nomothetic versus idiographic approaches, 2:684
- Individualized education programs (IEPs)
 - Individuals with Disabilities Education Act, 2:454–455
 - Information Referenced Testing, 2:460
- Individuals with Disabilities Education Act (IDEA), 2:454–456**
 - alternate assessment, 1:19
 - appropriate education, 2:455–456
 - child requirements, 2:454
 - Clinical Assessment of Behavior, 1:147
 - curriculum-based measurement, 1:208
 - due process protections, 2:455
 - IEPs, 2:454–455
 - least restrictive environment, 2:456
 - Section 504 versus, 3:876–877
 - state requirements, 2:454
- Individuals with Disabilities in Education Improvement Act (IDEIA), 3:840–842
- Induction
 - grounded theory and, 1:418, 420
 - idiographic approach, 2:684
 - null hypothesis significance testing, 2:695
- Inertia of vectors, 1:129
- Infantest Corporation, 1:351
- Inferential statistics, 2:457–460**
 - controversies over, 2:459–460
 - estimation, 2:457–459
 - hypothesis testing, 2:457–459
 - parameters, statistics, and probability, 2:457
 - Pearson product-moment correlation coefficient and, 2:753–754
 - samples and population, 2:457
 - Spearman's rho and, 3:929–932
 - statistical significance and, 3:964
- Inflated percent agreement statistics, 1:164–165
- Inflation of the alpha level, 1:103
- Information-error test, 1:54
- Information Referenced Testing (IRT), 1:114, 2:460–464**
- Information Systems Interaction Readiness Scales, 2:464–465**
- Informative prior distributions, 1:80
- Informed consent, 1:319, 2:465–468**
 - competence, 2:467–468
 - comprehension, 2:467
 - consent/approval, 2:468
 - disclosure, 2:466
 - questions about, 2:468
 - voluntary decision, 2:467
- Initial data. *See* Training data
- Inselberg, A., 2:727
- Institute of Mathematical Statistics, 1:27, 249
- Institute of Personality and Ability Testing, 3:917–918
- Instructional Topics in Educational Measurement Series*, 2:678
- Instrumental variable (IV) estimator, 2:470
- Instrumental variables, 2:468–472**
 - estimation, 2:470–472
 - example of, 2:469
 - properties of estimators, 2:471
 - specification tests, 2:472
- Intangible data, 1:220
- Integrity, in research, 1:324
- Intellectual maturity, 1:406
- Intelligence
 - achievement versus, 1:9
 - creativity and, 3:1010
 - definition, 1:2
 - Galton and, 1:383
 - Gf-Gc theory, 1:403–406

- job performance selection, 1:50
 theories of, 2:477
See also General intelligence (*g*); **Intelligence quotient (IQ)**
- Intelligence quotient (IQ), 2:473–475**
 definition, 2:474
 deviation IQ, 2:475
g and, 1:404
 history of, 2:473
 mental age, 2:474
 performance IQ, 2:757
 problems with, 2:474–475
- Intelligence tests, 2:476–479**
 achievement tests versus, 1:9, 2:478–479
 aptitude tests versus, 1:39–40
 CHC theory, 2:477–478
 contemporary, 2:478
 criticisms of, 2:478
 Culture Fair Intelligence Test, 1:204–205
 Fagan Test of Infant Intelligence, 1:351–352
 history of, 2:476–477
 intelligence theories, 2:477
 Lorge-Thorndike Intelligence Tests, 1:162
 Performance IQ, 2:757
 purpose, 1:2
 reaction-time theory, 2:478
 Universal Nonverbal Intelligence Test, 3:1026–1027
 verbal IQ, 3:1043–1044
 Wechsler Adult Intelligence Scale, 3:1047–1048
 Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III), 3:1049
- Intensity functions, 2:774
- Interactive statistical graphics, 1:414–416
- Interdependence
 covariance-level, 2:489
 ipsative measure, 2:488–489
 item-level, 2:489
- Internal consistency
 coefficient alpha, 1:155–157
 reliability and, 2:516–519, 3:834–835, 934, 944
 test-retest reliability and, 1:157
- Internal External Locus of Control Scale (I-E Scale), 2:479–480**
- Internal review board (IRB), 1:322, 325, 2:466, 468, 480–483**
 administrative review, 2:482
 authority, 2:482
 continuing review, 2:483
 full board review, 2:482–483
 history of, 2:480–481
 membership, 2:482
 operation of, 2:481–482
 purpose, 2:480
 review and approval basis, 2:481
 review types, 2:482–483
- Internal validity, 3:805–806
- International Assessment of Educational Progress (IAEP), 2:483–484**
- International Biometric Society, 1:27
- International Classification of Diseases, 2:429
- International Organization for Standardization, 3:863
- Interrater reliability, 2:484–486, 3:834**
 alternate assessment, 1:21
 Cohen's kappa, 1:164–165
 consensus, 2:484–485
 consistency, 2:484–485
 determination of, 1:155
 performance-based assessment, 2:759–760
- Interrupted time series designs, 3:806
- Interuniversity Consortium for Political and Social Research Guide to Resources and Services, 3:873
- Interval censoring, 1:125–126, 3:983–984
- Interval-contingent recording, 1:297
- Interval level of measurement, 2:486–487, 582**
- Interventions, 3:904
- Interviews
 face-to-face, 3:809
 in-depth, 1:252
 questionnaires, 3:809
 structured, 1:251
 telephone, 3:809
- Intraclass correlation, 2:435
- Intrinsic Bayes factors, 1:75
- Intron, 1:91
- Inverse binomial distributions, 2:773–774
- Inverse Fourier transform, 1:365
- Iowa Every-Pupil Tests, 2:487
- Iowa Test of Basic Skills (ITBS), 1:10, 162, 2:487**
- Iowa Tests of Educational Development (ITED), 1:162, 2:488**
- Ipsative measure, 2:488–489**
- Irrational numbers, estimation of, 2:628–629
- Irregular component, 3:1005–1006
- ISI Web of Knowledge, 3:872
- Isotropic scaling, 1:395
- Item analysis and elimination, 2:538–539
- Item and test bias, 2:489–493**
 differential item functioning, 2:490
 future issues, 2:492
 item bias, 2:490–491

- projective testing, 2:769
- self-report measures, 2:768, 3:877–878
- significance of, 2:489–490
- test bias, 2:491–492
- Item banks, 3:824–825
- Item characteristic curve, 2:493–494
- Item characteristic function, 2:493
- Item difficulty, 2:493, 517
- Item discrimination parameter, 2:493
- Item exposure rate, 1:172
- Item information function, 2:495
- Item-level interdependence, 2:489
- Item-person test of fit, 3:823
- Item pool, 1:173
- Item-response models, 2:615
- Item response theory (IRT), 2:493–496**
 - applications of, 2:495–496
 - assumptions of, 2:494
 - classical test theory, 1:142
 - differential item functioning, 2:490–491
 - item selection, 1:171–172
 - popular models, 2:493–494
 - properties, 2:494–495
 - scaling, 3:864–865
 - score equating/linking, 2:495–496
 - test construction, 2:495
- Item selection
 - adjectives, 3:879, 882
 - computerized adaptive testing, 1:171–173
 - item pool, 1:173
 - maximum information approach, 1:172
- Item-total score correlation, 2:538, 3:835
- Iterated function systems, 1:370–372
- Iteratively reweighted least squares (IRLS)
 - algorithm, 1:398
- Iterative methods, 2:532
- Iterative partitioning methods, 2:503
- IVEWARE for SAS, 2:664
- Jackknife methods, 1:169
- Jackofsky, E. F., 2:554
- Jackson, Douglas N., 1:71, 2:497, 640, 766**
- Jackson Personality Inventory–Revised (JPI-R), 2:498**
- Jackson Vocational Interest Survey (JVIS), 2:499**
- Jaeger, D., 2:471
- Jarrett, R. G., 1:317
- Jason, L. A., 1:56
- Jensen, A. R., 1:201, 204, 2:478
- Jensen difference, 1:278
- Jiang, F., 2:598
- Job analyses, 1:183
- Job lock, 2:426
- Job performance selection instruments
 - ability tests, 1:4
 - aptitude tests, 1:40–41
 - ASVAB, 1:45
 - intelligence tests, 1:50
 - Minnesota Clerical Test, 2:608
 - See also* Vocational instruments
- Joglekar, P., 1:244
- Johansson, Charles B., 1:118
- Johnson, J., 1:385
- Johnson, R. L., 2:486
- Johnson, V., 1:116
- Johnson, William D., 1:111
- Joint correspondence analysis, 2:656
- Joint Statistical Meetings, 1:27
- Jonckheere Terpstra test, 2:687
- Jöreskog, K. G., 2:733
- Journal de Mathématiques Pures et Appliquées*, 2:500
- Journal of Agricultural, Biological & Environmental Statistics*, 1:27
- Journal of the American Statistical Association*, 1:27, 2:499–500**
- Journal of Business & Economic Statistics*, 1:27
- Journal of Computational & Graphical Statistics*, 1:27
- Journal of Educational & Behavioral Statistics*, 1:27
- Journal of Educational Measurement*, 2:678
- Journal of Modern Applied Statistical Methods*, 2:500–501**
- Journal of Statistical Education*, 1:27
- Journal of Statistics Education*, 2:502**
- JPEG compression, 1:229–230
- Judges. *See* Evaluators; **Interrater reliability**; Raters
- Jung, Carl, 2:675
- Justice, in research, 1:324
- Just noticeable difference, 3:862
- $k = 1$ median, 2:593
- $k > 1$ medians, 2:593
- Kaiser criterion, 1:333
- Kaneko, Shuji, 3:872
- Kaplan-Meier estimator, 1:126, 3:983–985
- Karelitz, S., 1:94
- Karhunen-Loeve decomposition of Watanabe, 2:783
- Karlin, Samuel, 3:865
- Karvonen, M., 1:20
- Kassebaum, Nancy, 2:426
- Katz, J. A., 3:870, 873–876, 874
- Katz, S., 1:11

-
- Kaufman, Alan, 3:846
- Kaufman Ability Battery for Children, 2:478
- Kaufman Adolescent and Adult Intelligence Test, 3:846
- Kaufman Assessment Battery for Children (KABC-II)**, 2:507, 3:846
- Kazi-Aoual, F., 3:851
- Kelley, Truman, 1:2
- Kelly, K. P., 1:246
- Kendall, M. G., 1:358
- Kendall rank correlation**, 2:508–510
- example of, 2:508–509
 - generalization of, 2:510
 - notations and definition, 2:508
 - significance test, 2:509–510
- Kendall's tau statistic, 2:687
- Kennedy, D. T., 1:244
- Kennedy, Edward, 2:426
- Kenward, M. G., 1:263
- Kepler, Johannes, 3:896
- Kernel methods, 1:234, 3:979
- Kernel principal component analysis, 2:786
- Kerns, Robert D., 3:1050
- Keselman, Harvey J., 2:501
- Kiecolt, K. J., 3:870, 874
- Kim, J. J., 1:301
- Kim, Y. K., 3:1050
- Kinanthropometry, 1:37
- Kinetic Family Drawing Test**, 2:510–511
- Kingston Standardized Cognitive Assessment (KSCA-R)**, 2:512
- Kleiner, B., 2:631
- Klinke, S., 1:415
- Klockars, Alan, 2:501
- k-means cluster analysis**, 2:503–506
- data mining, 1:235
 - features, 2:504
 - randomization tests, 2:504
 - software packages, 2:503
- Knapp, Thomas R., 2:501
- Knowledge, achievement and, 1:8
- Knowledge discovery in databases (KDD), 1:170, 232
- Knowledge Factor, 1:114
- Known-groups method, 3:1033
- Kohonen networks, 1:234
- Kolmogorov distance, 2:513, 514, 515
- Kolmogorov-Smirnov goodness-of-fit test, 1:408
- Lilliefors test for normality and, 2:540
 - Shapiro-Wilk versus, 3:885
- Kolmogorov-Smirnov test for one sample**, 2:512–514
- Kolmogorov-Smirnov test for two samples**, 2:514–516, 687
- Korkin, A. N., 2:567
- Korth, B. A., 3:850
- KR-20 and KR-21**, 2:516–519, 3:835
- Kromrey, J. D., 1:344
- Kruskal, J. B., 3:864
- Kruskal-Wallis one-way analysis of variance**, 2:519–520, 687
- Kuder, Frederic, 2:521
- Kuder, G. Frederik, 1:155
- Kuder Career Search Schedule (KCSS), 2: 521
- Kuder Occupational Interest Survey**, 2:521
- Kuder-Richardson. *See* **KR-20 and KR-21**
- Kurtosis**, 2: 522–523, 674, 3:885
- L_1 , 1:166–167
- L_2 , 1:166–167
- Laboratory Behavioral Measures of Impulsivity**, 2:525–526
- Lag-1 autocorrelation coefficient, 1: 63–64
- Lag-2 autocorrelation coefficient, 1: 64
- Lagrange, Joseph Louis, 2:771
- Lagrangian multipliers, 1:308
- Laird, N., 1:310
- Lakhani, R., 3:971
- Landmarks, in anthropometry, 1:36, 37, 38
- Lane, C. J., 1:187
- Lane, David M., 1:135
- Langevin Metropolis-Hastings algorithm, 2:572
- Laplace, Pierre-Simon, 1:129, 2:690, 771
- Large-sample technique, 3:974
- Latent class analysis (LCA)**, 2:526–528
- Latent class models, 2:526–527
- Latent meaning, 3:999
- Latent means structures (LMS) analysis, 2:492
- Latent roots, 1:304
- Latent variable models. *See* **Mixture models**
- Latent variables, 1:79, 332, 341, 345, 3:973
- Latent vectors, 2:740–741
- Lattice data, 3:926
- Laugksch, R. C., 1:199
- Law of Averages, 2:528
- Law of comparative judgment, 3:1002
- Law of large numbers**, 2:528–529
- advanced definitions, 2:528
 - Bernoulli and, 1:87
 - Markov and, 2:568
 - misconceptions, 2:528
 - Poisson and, 2:772

- Law of Small Numbers, 2:528
Law School Admission Council (LSAC), 2:529
Law School Admissions Test (LSAT), 2:529–530
Lazar, M. D., 1:158
L.B. and J.B. v. Nebo School Dist., 2:456
Learning data. *See* Training data
Learning disabilities
 curriculum-based measurement, 1:208
 specific, 3:840–843
Learning with a teacher. *See* Supervised learning
Least restrictive environment, 2:456
Least significance difference (LSD) procedure, 1:359–360
Least squares, method of (LSM), 2:530–532
 drawbacks, 2:532
 examples of, 2:530–532
 history of, 2:530
 mean, 2:587
 popularity of, 2:530
Least variance ratio estimator, 2:471
Lebreton, J.-D., 3:851
Leclerc, George Louis, Comte de Buffon, 2:626, 628
Lee-Shin, Y., 3:1050
Left censoring, 1:125
Legendre, Adrien-Marie, 2:530, 691, 771
Lehmann, E. L., 3:1026
Lehrbuch der Anthropologie (Martin), 1:36
Leibniz, Gottfried Wilhelm, 1:86, 2:500, 530
Leighton, J. P., 2:439
Lempel, Abraham, 1:227
Leno, Jay, 2:719
Leong, F. T. L., 1:121
Leptokurtic distribution, 2: 522
Lesaffre, E., 1:263
Leung, M.-Y., 3:867
Levant, Ronald F., 2:563
Level of significance, 3:889–891, 966, 1019
Levene's test, 2:443
Levin, J. R., 1:35
Lewis, A., 3:988
Lewis, J. G., 1:15
Liang, K.-Y., 1:400
Life Values Inventory, 2:532–533
Lightner Witmer Award, 1:417
Ligon, Ernest M., 1:174
Likelihood, 2:457
Likelihood function, 1:80
Likelihood ratio test, 2:533–536
 calculating *p* values, 2:534–535
 power properties, 2:534
 software packages, 2:535
Likert, Rensis, 1:53, 2:536, 3:843
Likert scaling, 2:536–539
 attitude tests, 1:53
 constructing Likert scale, 2:537–539
 evaluation of, 2:539
 history and development, 2:536–537
 KR-20, 2:519
 semantic differential scale versus, 3:880
Lilliefors, H. W., 2:514, 540
Lilliefors test for normality, 2:540–544
 notation, 2:540–541
 numerical approximation, 2:542–544
 numerical example, 2:541–542
Limited information maximum likelihood, 2:470–471
Limiting quality level (LQL), 1:6
Linearity, 1:366
Linear mixed-effects model, 2:612–614
Linear models, 1:34
Linear programming, 1:172
Linear regression, 2:544–550, 3:829–830
 coefficients, 2:549
 correlations, 2:546–548
 multiple, 2:546
 simple, 2:545–546
 weights, 2:548–549
 See also Ordinary least squares (OLS) regression
Linear regression analysis, 1:168
Linear regression through the origin, 2:431
Line chart, 2:544
Line of best fit, 3:829
Linex error loss, 1:237–238
Linked highlighting, 1:415–416
Link function, 2:774
Linn, R. L., 2:758, 759
Liouville, Joseph, 2:500, 568, 772
Lipsey, Mark W., 2:596
LISREL, 1:59
List of American Doctoral Dissertations Printed, 1:25
Little, Brian R., 2:762, 764
Little, R. J. A., 1:261, 263
Ljung-Box test, 1: 64
Loading matrix, 2:741
Local item independence, 2:494
Local principal component analysis, 2:785–786
Location parameter, 2:690
Locus of control. *See* **Internal External Locus of Control Scale**
Logarithmic cooling schedule, 3:898–899
Logarithmic series distribution, 2:773
Logical zooming, 1:416

- Logistic curve, 1:185
- Logistic regression analysis**, 2:550–554
- categorical outcomes, 3:830
 - conditional, 3:833
 - logit transformation, 2:551
 - multiple, 2:553–554
 - odds, 2:551
 - ordinary versus, 2:550–551
 - relative risk, 3:833
 - simple, 2:551–553
- Logit, 2:551, 3:821
- Loglinear analysis**, 2:554–558
- Log-rank test, 1:126, 3:985–986
- Lombroso, Cesare, 1:36–37
- Longitudinal/repeated measures data**, 2:558–561
- advantages, 3:836
 - attrition bias, 1:57
 - Bowker procedure, 1:107–111
 - complex, 2:559
 - Diggle-Kenward Model for Dropout, 1:261–264
 - disadvantages, 3:836
 - generalized estimating equations, 1:397–403
 - limitations of, 2:561
 - McNemar Test for Significance of Changes, 2:576–580
 - mixed models, 2:612
 - parameter invariance, 2:735
 - simple, 2:558–560
 - SPSS calculation of, 2:560
- Long-term retrieval, 1:3
- Long-term storage and retrieval, 1:405
- Looney, S. W., 2:674–675
- Lopez, M. N., 1:158
- Lord, F. M., 1:142
- Lord, Frederic, 1:97
- Lord, M. F., 1:171–172
- Lorge-Thorndike Intelligence Tests, 1:162
- Loss function, 1:237–238
- Lossless data compression, 1:226
- Lossy data compression, 1:226
- Low-pass filter, 1:368
- L* test, 2:719–722
- Luce, R. D., 2:583
- Ludwig of Bavaria, 2:707
- Luria, A. R., 2:561
- Luria Nebraska Neuropsychological Battery**, 2:561–562
- Lutz, M. N., 2:454
- Lutzker, J. R., 1:127, 3:819
- Lynam, D. R., 1:258
- Lytle, L. A., 2:616
- LZ77, 1:227
- LZW, 1:227–228
- Machine learning, 1:233, 3:978
- MacKeith Press, 2:680
- MacKinnon, D. P., 3:996, 998
- MacQueen, James B., 2:503
- Mager, R. F., 1:195
- Magnitude estimation, 3:863–864
- Magnitude production, 3:863
- Magnitude scaling, 3:863–864
- Mahalanobis distance, 1:209–210, 282–283
- Main effects, 1:33–34
- Majovski, Lawrence, 2:562
- Malebranche, Nicolas, 1:86
- Male Role Norms Inventory**, 2:563–564
- Malthus, Thomas**, 2:564
- Mandelbrot, Benoit, 1:373
- Manhattan Project, 2:627
- Manifest meaning, 3:999
- Mann-Whitney U test (Wilcoxon Rank-Sum test)**, 2:565–567, 687
- MANOVA. *See* **Multivariate analysis of variance (MANOVA)**
- Mantle, Mickey, 2:451
- Many-Facets Model of Rasch (MFMR), 3:824
- Marascuilo, L. A., 1:153
- Mardia's measures of skewness and kurtosis, 2:674
- Marginal significance, 2:697
- Maritz, J. S., 1:317
- Maritz-Jarrett estimate, 2:593
- Markov, Andrei Andreevich**, 2:567–568
- Markov, Andrei, Jr., 2:567
- Markov chain Monte Carlo methods (MCMC)**, 2:568–573
- ergodic theorem, 2:569–570
 - Gibbs sampler, 2:572
 - issues in, 2:573
 - Metropolis-Hastings algorithm, 2:571–572
 - Metropolis sampler, 2:570–571
 - multiple imputation for missing data, 2:664
 - rejection sampling, 2:569
 - simulated annealing, 2:570–571
- Markov chains, 2:568, 3:897
- See also* **Markov chain Monte Carlo methods**
- Markov field, 2:568
- Markov process, 2:568
- Markov property, 2:568
- Marlowe-Crowe Social Desirability Scale, 2:479
- Marsaglia, George, 2:501

- Marsella, A. J., 3:1005
Martens, H., 2:744
Martin, Rudolf, 1:36
Masking, 1:229
Mason, M., 1:58
Massey, F. J., Jr., 1:294
Matched case control study, 3:833
Matched pairs, 3:828
Matching
 exact, 2:527, 3:827–828
 statistical, 3:828
Matrices
 determinant of, 1:307–308
 DISTATIS, 1:286–289
 full-rank, 1:308
 negative semi-definite, 1:307
 positive definite, 1:307
 positive semi-definite, 1:306–307
 rank of, 1:308
 STATIS, 3:957–962
 trace of, 1:307
Matrix Analogies Test, 2:573–574
Matrix correlation coefficients, 3:849
Matrix operations, 2:574–575
Mauchly's test, 2:560
Maximization. *See* **EM algorithm**
Maximum information (MI) approach, 1:172
Maximum likelihood estimate (MLE), 2:576
 EM algorithm, 1:310–313
 mixed-effects models, 2:614
 mixture models, 2:617–619
 mixtures-of-experts models, 2:622
 structural equation modeling, 3:974–975
Maximum likelihood method, 1:97, 334, 2:575–576
May, Warren L., 1:111
McArdle, J. J., 2:475
McCrae, Robert, 2:679
McDaniel, F., II, 2:486
McDonald, R. P., 2:733
McGill Pain Questionnaire, 3:1050
McGrew, Kevin, 1:404
McKinley, J. C., 2:609
McNemar, Quinn, 1:107, 2:576, 3:980
McNemar Test for Significance of Changes, 2:576–580
 application of, 2:578–579
 Bowker procedure and, 1:107–111
 Cochran Q test and, 1:154
 confidence interval estimate, 2:579–580
 development of, 2:577
 directional versus nondirectional, 2:579
 matched/paired differences, 2:579
 testing for significance, 2:577–578
McSweeney, M., 1:153
Mean, 2:580–581
 advantages, 2:590–591
 arithmetic, 1:43–44, 65, 2:586–587
 average deviation, 1:66
 centroid and, 1:128
 comparisons involving, 2:590–591
 disadvantages, 2:591
 geometric, 1: 65, 2:587–588, 591
 grand, 1:411–412
 harmonic, 1: 65, 2:425–426, 588
 measure of central tendency, 2:586–588
 midmean, 1: 65
 sample versus population, 2:732
 trimean, 2:587, 591
 trimmed, 1: 65, 2:587, 591
 weighted, 2:587
Mean deviation, 1:66
Mean difference, 3:964
Mean exponential family distributions, 2:775
Meaningfulness of relationship, 3:1030–1031
Mean squared error, 1:238, 2:715
Measurement, 2:582–585
 historical perspectives, 2:582–583
 levels of, 2:582–583, 3:826, 863
 modern perspectives, 2:583–584
 statistics versus, 2:582, 583
 See also **Anthropometry**
Measurement error, 2:585
 Cronbach's study of, 1:204
 difference scores and, 1:259
 random versus systematic, 2:585
 reducing, 2:585
 systematic versus unsystematic, 1:141, 2:730, 3:944, 992–993
 See also Error
Measurement models, 3:973
Measure of dispersion, 3:882–883
Measures of central tendency, 2:586–591
 arithmetic mean, 1:43–44
 average, 1: 65–66
 comparisons of, 2:590–591
 mean, 2:586–588
 median, 2:588–589
 mode, 2:586
 semi-interquartile range, 3:882–883
 software packages, 2:589–590
 See also Central tendency

- Measures of health-related quality of life (HRQOL), 3:804
- Mecklin, C. J., 2:674
- Median, 2:591–592**
 advantages/disadvantages, 2:591
 average and, 1:65
 average deviation, 1:66
 comparisons involving, 2:590–591
 Delphi technique, 1:242
 estimates of the population, 1:316–318
 measure of central tendency, 2:588–589
 percentile and, 2:756
 population, 2:593
 sample, 1:316–318, 2:593
 semi-interquartile range, 3:882–883
 testing for, 2:592–594
- Median test, 2:592–594**
- Mediator variables, 2:624, 3:974, 995–998
- Medical informatics, 1:90
- Medicine, anthropometry and, 1:37
- Meehl, Paul, 3:1034
- Memory
 Immediate and Delayed Memory Tasks, 2:452–453
 memory-based reasoning, 1:234–235
 short-term, 1:3, 405, 3:951
- Memory-based reasoning, 1:234–235
- Memo writing, 1:419
- Mental age, 2:474, 476, 3:950–951
- Mental measurements, 1:115
- Mental Measurements Yearbook*, 1:115, 2:478
- Mercer, S., 3:988
- Meredith, W., 2:733
- Mesmer, Franz Anton, 2:709
- Messick, S., 1:182, 2:757, 3:1034
- Meta-analysis, 2:595–598**
 criticisms of, 2:595
 data analysis/interpretation, 2:596
 data collection, 2:596
 data evaluation, 2:596
 definition, 1:353
 drawbacks of, 1:353
 effect size estimation, 2:596–597
 effect size testing, 2:597
 face validity of, 1:339–340
 presentation of findings, 2:596
 publication bias, 2:597
 research hypothesis, 2:595
 secondary data analysis versus, 3:869–870
 stages of, 2:595–596
- Method of equal-appearing intervals, 3:1002–1003
- Method of moments estimation, 1:390–394, 2:617
- Method of paired comparisons, 3:1003–1004
- Method of successive intervals, 3:1004
- Method of summated ratings. *See* **Likert scaling**
- Metric multidimensional scaling (MDS), 2:598–604**
 example of, 2:598
 nonmetric data analysis, 2:603–604
 proof, 2:604
 singular value decomposition and, 3:907
 supplementary elements, 2:601–603
 transforming distance to cross-product matrix, 2:599–601
- Metropolis, N., 3:896
- Metropolis-Hastings algorithm, 2:571–572
- Metropolis simulation algorithm, 2:570–571, 3:896–897
- MICE software package, 2:664
- Michell, J., 2:583–584
- Microarray technology, 1:90
- Midmean, 1:65
- Midpoints, 3:810–811
- Milgram Obedience Studies, 2:466
- Mill, John Stuart, 1:124
- Miller, J. D., 1:258
- Miller, M. D., 2:758
- Millon, Theodore, 2:606, 768
- Millon Adolescent Clinical Inventory (MACI), 2:606
- Millon Adolescent Personality Inventory (MAPI), 2:606
- Millon Behavioral Health Inventory (MBHI), 2:605
- Millon Behavioral Medicine Diagnostic (MBMD), 2:605**
- Millon Clinical Multiaxial Inventory-III (MCMI-III), 2:606, 768, 769**
- Millon Inventory of Personality Styles-Revised (MIPS-Revised), 2:606
- Millon Pre-Adolescent Clinical Inventory (M-PACI), 2:606
- Minard, Charles Joseph, 1:412
- Mind Garden, 1:352, 3:920
- Minimax, 1:239
- Mini Mental State Examination (MMSE), 2:512
- Minimum average variance estimation (MAVE), 1:266
- Minimum residual method of factor extraction, 1:174
- Minitab, 2:633–637
- Minkowski's distances, 1:283
- Minnesota Clerical Test, 2:607**
- Minnesota Multiphasic Personality Inventory (MMPI), 2:607–608**
 Comrey Personality Scales and, 1:174

- Millon Clinical Multiaxial Inventory-III and, 2:606
 Personality Assessment Inventory and, 2:766
 psychopathology assessment, 2:769
 Minnesota Vocational Test for Clerical Workers, 2:608
 Misinformation, 2:460
 Miss, 2:697–698
 Missing at random, nonresponse process, 1:261–263
 Missing completely at random, nonresponse process, 1:261–262
Missing data method, 2:608–612
 development of, 2:608–609
 multiple imputation for missing data, 2:663–666
 principal-component method, 2:610–611
 regression model, 2:609–610
 weights, 3:981
 See also **Censored data**
 Mixed ANOVA, 1:33
 Mixed-effects models. See **Mixed models**
 Mixed-format scales, 3:845
Mixed models, 2:612–616
 hierarchical linear modeling, 2:432–435
 repeated measures data, 3:837
Mixture models, 2:616–619
 discriminant analysis, 1:269
 finite mixture distributions, 2:616–617
 latent class analysis, 2:526
 latent variables, 2:619–620
 maximum likelihood method, 2:617–618
 mixtures-of-experts models, 2:620
 unknown number of components, 2:618–619
Mixtures of experts, 2:619–622
 generalized linear models, 2:620
 mixture models, 2:620
 Modal a posteriori estimate, 1:80, 2:781
Mode, 2:623
 advantages/disadvantages, 2:591
 average and, 1:65
 comparisons involving, 2:590–591
 measure of central tendency, 2:586
 Model comparison
 Bayes factors, 1:74–76
 Bayesian information criterion, 1:76–79
 Model fitting, 1:334
 Model selection
 criteria, 1:15, 78, 2:748
 process, 2:558
 Moderated multiple regression, 2:625–626
Moderator variable, 2:624–626
 analysis of variance based test, 2:624–625
 detecting, 2:624
 homogeneity of correlation coefficients, 2:626
 interactions in multiple regression, 1:47–48
 issues involving, 2:626
 mediator variable versus, 2:624
 moderated multiple regression based test, 2:625–626
 Modified double-exponential distribution, 1:317–318
 Mohr, W. K., 2:454
 Moivre, Abraham de, 1:129, 2:690, 772
 Molecular human biology, 1:91, 3:866–867
 Molenberghs, G., 1:263
 Molin, P., 2:540, 542
 Moment conditions, 1:390
 Moment functions, 1:390
Monatliche Auszug, 2:500
Monographs of the Society for Research in Child Development, 3:922
 Monotonic effects, 2:719–720
 Monotonicity coefficient, 3:849
 Monotrait heteromethod diagonals, 2:666–668
 Monotrait monomethod diagonals, 2:667–668
Monte Carlo methods, 2:626–630
 accuracy of, 2:629
 applications of, 2:630
 Bernoulli trials, 1:87
 bioinformatics, 1:92
 chance, 1:131
 congruence coefficient, 3:850
 estimating area of irregular figure, 2:628
 estimating area of regular figure, 2:627–628
 estimating irrational numbers, 2:628
 family of tests, 1:104–105
 hypothesis testing, 2:447
 Lilliefors test, 2:540
 Markov chain Monte Carlo, 2:568–573
 overview of, 2:568–569
 programming environment, 2:627
 pseudo random number generators, 2:627
 random number generation, 1:169
 simulating die tosses, 2:627
 software packages, 2:629
 static models, 3:901
 statistical inference, 1:170
 study versus simulation, 2:629–630
 Monte Carlo simulation, 2:629–630
 Monte Carlo study, 2:629–630
 Montgomery-Asberg Depression Rating Scale, 1:121
 Mood's test, 2:594
 Moore, J. B., 1:38
 Morgan, Christiana, 3:1000
 Morgenstern, Oskar, 1:237

- Morse, H., 3:1005
Morse, Samuel, 1:386
Morton, Samuel George, 1:36
Mosaic plots, 2:631–633
Mostert, Mark, 1:339
Motorola, 3:913–914
Mountin, Joseph W., 1:126
Moving average, 2:633–638
 applications of, 2:637–638
 calculating, 2:633–637
 software packages, 2:633–637
 uses, 2:633
MP3 encoder, 1:229
MPEG2, 1:230–231
Mueller, D. J., 2:539
Multiaptitude tests, 1:2, 41
Multicollinearity, 2:638–640
 detecting, 2:639–640
 effects of, 2:639
 remedial measures, 2:640
Multicollinear matrix, 1:308
Multidimensional Aptitude Battery (MAB-II), 2:640–641
Multidimensional scaling, 1:265
Multidimensional Stimulus Fluency Measure, 3:1012
Multi-Health Systems, 1:121, 2:442, 3:977
Multilayer perceptrons, 1:234
Multilevel growth modeling. *See* **Longitudinal/repeated measures data**
Multilevel modeling. *See* **Hierarchical linear modeling**
Multimodal distribution, 2:623
Multinomial logistic regression, 3:830
Multiple Affect Adjective Checklist–Revised (MAACL-R), 2:641–642
 reliability, 2:641
 validity, 2:641–642
Multiple-baseline designs, 3:906–907
Multiple-choice items, 2:642–644
 advantages/disadvantages, 2:644
 completion items versus, 1:167
 controversy over, 2:642
 forms, 2:642–643
 history of, 2:642
 Information Referenced Testing, 2:460–463
 number of options in, 2:643
 writing guidelines, 2:643
Multiple comparisons, 2:644–647
 handling, 2:646
 performing, 2:646–647
 power, 2:646
 special treatment necessary for, 2:644–646
Multiple comparison tests, 1:295
Multiple correlation coefficient, 2:648–651
 estimating population correlation, 2:649
 multiple regression framework, 2:648–651
 nonorthogonal predictors, 2:650–651
 orthogonal predictors, 2:649–650
 significance test, 2:649
Multiple correspondence analysis (MCA), 2:651–656
 alternatives, 2:656
 eigenvalue correction for, 2:653–654
 example of, 2:652, 654–656
 interpretation, 2:654–656
 notations, 2:652–653
 principal component analysis and, 2:785
 uses, 2:651
Multiple factor analysis (MFA), 2:657–663
 analyzing between-study structure, 2:660–663
 example of, 2:657–663
 global matrix, 2:658–659
 notations, 2:658
 partial analyses, 2:659–660
 performance of, 2:657
Multiple imputation for missing data, 2:663–666
 assumptions of, 2:665
 extensions, 2:666
 illustration of, 2:664–665
 software packages, 2:665
Multiple indicator, multiple cause modeling, 2:492
Multiple intelligences theory, 1:404
Multiple-item direct attitude measures, 1:53–54
Multiple linear regression, 2:546, 736
Multiple regression analysis
 advantages, 3:830
 chance and, 1:132
 missing data method, 2:609–610
 multicollinearity, 2:638–640
 multiple correlation coefficient, 2:648–651
Multiplication Rule, 3:1012
Multistate life table method, 1:11
Multitrait-multimethod matrix, 1:179
Multitrait multimethod matrix and construct validity (MTMM), 2:666–669
 goal of, 2:666
 interpretation, 2:667–668
 limitations of, 2:668–669
 organization of, 2:666–667
 trait and method selection, 2:666

Multivariate analysis of variance (MANOVA), 2:669–672

- example of, 2:670–671
- factorial, 2:669
- one-way, 2:669
- repeated measures design, 3:837
- shortcomings of, 2:672
- tests of statistical significance, 2:670–672

Multivariate distance, 1:209–210

Multivariate normal distribution, 2:673–675

- failure to test for, 2:674–675
- goodness-of-fit tests, 2:674
- normal curve, 2:695
- properties, 2:673–674
- significance of, 2:674

Multiway frequency analysis, 2:554

Mumford, K. R., 1:344

Mundfrom, D. J., 2:674

Murphy, Gardner, 2:536

Murray, Charles, 1:45

Murray, D. M., 2:616

Murray, H. A., 1:299

Murray, Henry A., 3:1000

Mutations, 1:234

Myers, Isabel, 2:675

Myers-Briggs Type Indicator (MBTI), 2:675–676, 770

Nagelkerke R Square, 2:554

Naisbitt, John, 1:91

Naive effect estimate, 2:705

Napier, H. Al, 1:135

Napoleon's March to Moscow, 1:412

Narrowing of curriculum, 2:438

Nathan, L. E., 3:870, 874

National Academies of Science, 3:922

National Assessment of Educational Progress (NAEP), 2:483

National Association of Teachers of Education Measurement, 2:677

National Bureau of Standards, Statistical Engineering Laboratory, 1:248

National Career Assessment Services, 2:521

National Center for Education Statistics, 1:150, 3:873

National Center for Health Statistics, 1:150, 3:828

National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research, 1:322, 2:481

National Commission on Education, 2:437

National Council of State Boards of Nursing, 1:171

National Council of Teachers of Mathematics, 1:10

National Council on Measurement in Education, 1:183

National Council on Measurement in Education (NCME), 2:677–678

- achievement awards, 2:678
- annual convention, 2:678
- formation, 2:677
- publications, 2:678
- purpose, 2:677

Standards for Educational and Psychological Testing, 3:947

National Council on Measurements Used in Education, 2:677

National Death Index, 3:828

National Health and Nutrition Examination Survey, 2:663

National Institute of Neurological and Communication Disorders and Stroke, 2:512

National Institutes of Health (NIH), 2:481

National Research Act (1974), 1:322, 2:480–481

National Research Council, 3:922

National Science Board, 2:679

National Science Foundation (NSF), 2:678–679

- American Statistical Association, 1:27
- grants and merit review, 2:679
- influence of, 2:679
- organizational structure, 2:678–679
- purpose, 2:678

A Nation at Risk (National Commission on Education), 2:437

NATO, 1:114, 2:463

Natural selection, 1:217, 234

Naus, Joseph, 3:865

Navy General Classification Test, 1:45

NCE scores, 3:954

NCS Assessments, 2:605

Nearest neighbor methods, 1:234–235

Negative predictive value (NPV), 1:256

Negative semi-definite matrices, 1:307

Negative suppression, 3:980

Neighborhood structure, 3:926

Nelder, J. A., 1:398

Neonatal Behavioral Assessment Scale (NBAS), 2:680–681**NEO Personality Inventory, 2:679–680, 770**

Neuman, Lawrence, 3:873

Neural networks

- artificial, 1:46–47
- data mining, 1:234

Neurons, 1:46

Neuropsychology test battery, 2:561–562

Nevo, Baruch, 1:338–339

- Newman-Keuls test**, 2:681–683
 example of, 2:682–683
 problems with, 2:681
- Newton, Isaac, 1:86, 2:530, 3:896
- Newton-Cotes integration methods, 3:895
- New York Psychiatric Institute Biometrics Research Lab, 3:976
- Neyman, J., 1:238
- Neyman allocation, 3:969
- Neyman-Pearson lemma, 1:238, 2:534
- Nezworski, M. T., 3:848
- Nightingale, Florence, 1:27, 412–414
- No Child Left Behind (NCLB) Act, 1:19–20, 2:437, 3:990
- N of 1 designs. *See* **Single-subject designs**
- Noise, 2:585, 3:886–887, 918–919
- Nominal Group Technique, 1:241, 242, 245
- Nominalism, 2:708
- Nominal level of measurement**, 2:683–684
 definition, 2:582
 interval versus, 2:486–487
- Nominal variables, 1:122
- Nomothetic versus idiographic**, 2:684
 personal projects analysis, 2:763–765
 Rorschach Inkblot Test, 3:848
- Noncompensatory model, 1:163
- Nondirectional alternative hypothesis, 1:23–24
- Nondirectional hypothesis, 2:459
- Nondirectional test, 2:712–713
- Nonequivalent control group designs, 3:806
- Nonexperimental studies, and mediating effects, 3:995–997
- Nonhomogeneous Poisson process, 2:774
- Noninformative prior distributions, 1:80
- Nonlinear regression, 1:214–215
- Nonmaleficence, in research, 1:323–324
- Nonparametric regression, 2:688, 3:919
- Nonparametric statistics**, 2:685–689
 advantages, 2:685
 one-sample location problem, 2:686–687
 parametric versus, 1:407, 2:685
 sample sizes, 1:407
 software packages, 2:688
 standard methods, 2:685–688
 two-sample problems, 2:687
- Nonprobability sampling**, 2:689–690, 787, 3:812
- Nonrandomized experiments. *See* **Quasi-experimental method**
- Nonresponse options, 3:811
- Nonresponse to surveys
 accounting for, 3:981
 random versus nonrandom, 1:261–263
- Nonsense correlation, 3:938
- Nonspuriousness, 3:937
- Nonuniform differential item functioning, 2:490
- No-predictors two-level hierarchical linear model, 2:434
- Normal approximations, 1:88, 90
- Normal cumulative distribution function, 2:710
- Normal curve**, 2:690–695
 areas beneath, table of, 3:1062–1063
 cautions about, 2:693–695
 example of, 2:692–693
 history of, 2:690
 properties, 2:691–692
 significance of, 2:693
 software packages, 2:694–695
- Normal distribution, 1:385
 bell curve, 1:206
 multivariate, 2:673–675
- Normality
 Lilliefors test for, 2:540
 Q-Q plot for assessing, 3:802–804
 Shapiro-Wilk test for, 3:883–885
- Normal probability plot, 3:802
- Normed fit index, 3:975
- Norm-referenced interpretations, 1:9
- Norm-referenced tests
 criterion-referenced versus, 1:195
 stanines and, 3:953
 validity, 1:200
- Norms, 2:474
- NORM software package, 2:664
- Norris, S. P., 2:439
- North Atlantic Treaty Organization (NATO), 1:114, 2:463
- North Carolina State University, Department of Statistics, 2:502
- Notched box plot, 1:112
- Nouvelle Correspondance Mathématique*, 2:500
- Novick, M. R., 1:142
- Nuisance variables, 1:29
 See also Covariates
- Null hypothesis
 alternative hypothesis versus, 1:23
 correlation coefficient values for rejecting, table of, 3:1070
 critical value and, 1:202–203
 hypothesis testing, 2:446, 711–712
 inferential statistics, 2:458
 number of tests and rejection of, 1:103–104
 t values for rejecting, table of, 3:1064–1065

- Type II error, **3:1019–1023**
See also **Null hypothesis significance testing**
- Null hypothesis significance testing (NHST), 2:695–698**
 decision making, **2:697–698**
 full-suite analysis, **2:696–697**
 induction, **2:695**
 power, **2:697–698**
 sampling distribution, **3:860–861**
 statistical significance, **3:963**
 Type I error, **3:1019**
See also Null hypothesis
- Null model, **2:434**
- Numerical analysis, **1:168–169**
 evaluation of special functions, **1:169**
 numerical linear algebra, **1:168**
 numerical optimization, **1:168–169**
 random number generation, **1:169**
- Numerical linear algebra, **1:168**
- Numerical optimization, **1:168–169**
- Nunnally, J. C., **1:51**
- Nuremberg Code, **1:322, 2:466, 480, 481**
- Nuremberg Trials, **1:322, 2:480**
- Nyquist frequency, **1:368**
- Objective-referenced tests, **1:197**
- Object specific scale, **2:539**
- Oblique rotations, **1:335**
- O'Brien, Ralph G., **2:701**
- O'Brien test for homogeneity of variance, 2:701–704**
 example of, **2:702–704**
 motivation and method, **2:701–702**
- Obsequiousness Bias, **2:430**
- Observational studies, 2:704–706**
 ANCOVA and, **1:29**
 bias, **2:705–706**
 descriptive research, **1:251–252**
 design and practicality, **2:704–705**
 experimental versus, **2:704–705**
 interpretation, **2:705**
 participant observation, **1:252**
See also **Quasi-experimental method**
- Observations, and curse of dimensionality, **1:209–211**
- Observer (APS), **1:49**
- Occupational Information Network (O*NET)
 Ability Profiler, **1:4**
- Occupational instruments. *See* Vocational instruments
- Occupational multiaptitude tests, **1:41**
- Ockham's razor, 2:707–709**
 example of, **2:709**
 science, **2:708–709**
 theological/philosophical background, **2:707–708**
- Odd-even estimate of internal consistency
 reliability, **2:517**
- Odds, **2:551**
- Odds ratio, **3:833**
- Ogburn, William F., **2:500**
- Ogive, 1:205–206, 2:710–711**
- Oh, S., **1:158**
- Oldman, D., **1:133**
- Olejnik, S., **1:301**
- Olkin, Ingram, **2:595, 597**
- Olmstead-Schafer, M., **1:244**
- Olson, D. H., **1:72–73**
- Olson, J. R., **2:430**
- One- and two-tailed tests, 1:23–24, 2:711–713**
- One-factor ANCOVA, **1:31**
- One-factor ANOVA, **1:30–31**
- One-group posttest-only designs, **3:806**
- One-parameter model, **3:821**
- One-way analysis of variance, 1:32, 2:713–716**
 assumptions of, **2:716**
 example of, **2:714–716**
 follow-up tests, **2:716**
 Newman-Keuls test, **2:681–683**
 Peritz procedure, **2:760–762**
 repeated measures design, **2:716**
- One-way independent ANOVA. *See* Between-groups ANOVA
- One-way multivariate analysis of variance, **2:669**
- One-way random effects analysis of variance model, **2:434**
- One-way repeated measures ANOVA, **1:33, 2:558–560, 713–714**
- One-way repeated measures design, **2:558–560**
- Open coding, **1:419**
- Open-ended questions
 essays, **1:315**
 questionnaires, **3:809**
 self-report measures, **3:877**
- Open source movement, **1:170**
- Opinion, definition of, **3:1002**
- Opinion polls, **1:251**
- Optimal allocation, **3:969–970**
- Optimal GMM estimator, **1:391**
- Optimal scaling, **2:651**
- Optimization, eigendecomposition and, **1:308**
- Order effects, **3:810, 811**
- Order statistics, **1:291**

Ordinal level of measurement, 2:716–717

- definition, 2:582
- interval versus, 2:486–487
- pairwise comparisons, 2:726–727

Ordinal logistic regression, 3:830

Ordinal variables, 1:122

Ordinary least squares (OLS) estimator, 2:468–472

Ordinary least squares (OLS) regression, 1:265,
2:431–432, 530–531, 3:829–830

See also **Linear regression**

Orthogonal matrix, 1:306

Orthogonal predictors in regression, 2:717–718

Orthogonal rotations, 1:335

Orwin, Robert G., 2:596

Osgood, Charles, 1:54, 3:878

Osgood, D. W., 2:688

Osgood's model of communication, 2:452

Ostoff, Johanna, 1:385

Otherwise qualified individuals, 3:876–877

O'Toole, A. J., 1:284, 2:598

Ott, Jurg, 3:867

Outcome variable, 2:442

Outliers

- arithmetic mean, 1:44
- box and whisker plots, 1:111
- coefficient alpha, 1:160
- coefficient of correlation, 1:158
- Dixon test for, 1:291–293
- hypothesis test for, 1:291
- least squares method, 2:532
- methods for handling, 1:291

Outstanding Statistical Application Award, 1:27

Overachievement, 1:42–43

Overdispersion in data, 2:773

Overidentification, instrumental variables and,
2:470, 472

Overstriking, 3:977

Overt bias, 2:705–706

Page, Ellis B., 2:719

Page's L Test, 2:719–722

- application, 2:720–722
- development of, 2:719–720
- Friedman test versus, 2:722
- power, 2:722

Pain measures, 3:1050–1051

Paired replicates data, 2:686–687

Paired samples *t* test (dependent samples *t* test), 2:723–726

- advantages, 2:724–726

- assumptions of, 2:723
- computations, 2:724–725
- example of, 2:723–725
- SPSS calculation of, 2:724–725
- uses, 2:723

Pairwise comparisons, 2:726–727, 778

Pan, W., 1:400

Panel data, 1:397

Parallel analysis, 1:334

Parallel coordinate plots, 2:727–729

- advantages, 2:728
- disadvantages, 2:728
- example of, 2:728–729
- properties, 2:728

Parallel forms reliability, 2:730–731, 3:993

Parallel tests, 1:141–142

Parameter, 2:731–732

- definition, 2:457, 731–732, 3:963
- inferential statistics, 2:458
- maximum likelihood method, 2:575–576
- population and, 2:732
- random sampling and estimation of, 3:817–818
- statistic versus, 2:732

Parameter-expanded EM algorithm, 1:313

Parameter invariance, 2:733–735

- across examinee subgroups, 2:734–735
- across items, 2:733–734
- across time points, 2:735
- bias versus impact, 2:734
- definition, 2:733
- generalizability of inferences and, 2:733
- item- and test-level effects, 2:735
- mathematical formalization of, 2:733
- methodological approaches, 2:733
- uniform, nonuniform, and crossing effects, 2:734–735

Parameter space, 1:237

Parametric statistics

- ANCOVA and, 1:32
- distributional assumptions, 1:407
- nonparametric versus, 1:407, 2:685

Parsimony, 2:708–709, 3:1037–1041

Part and partial correlations, 2:736–739

- F* and *t* tests, 2:738
- increment in explained variance, 2:736–737
- multiple regression framework, 2:736
- prediction from residual, 2:737–738

Partial credit model (PCM) of Rasch, 3:822–823

Partial least square regression (PLS), 2:740–744

- algorithm for, 2:741
- covariance, 2:741

- decomposition of predictors and dependent variables, 2:740–741
- example of, 2:742–744
- goal of, 2:740
- prediction of dependent variables, 2:741–742
- relationship with other techniques, 2:744
- singular value decomposition, 2:741
- software packages, 2:744
- Partial least squares method, 2:530
- Partial likelihood, 3:987
- Partial mediation, 3:995, 997
- Participant observation, 1:252
- Partition function, 3:896
- Partitioning data, 1:169
- Pascal, Blaise**, 1:129, 2:744–745
- Paterson, D. G., 1:2
- Path analysis**, 2:745–748
 - assumptions of, 2:747
 - causal analysis and, 1:125
 - comparing alternative models, 2:748
 - conditions for estimability, 2:748
 - diagrams, 2:747–748
 - examples of, 2:745–746
 - regression, 2:746
 - software packages, 2:748
 - structural equation modeling versus, 2:747, 3:973
 - tests of mediating effects, 3:996
- Path coefficients, 2:747
- Path diagram, 1:342, 2:747–748
- Patil, K. D., 1:152
- $p_{\text{calculated}}$, 3:889–891
- p_{critical} , 3:889–891
- Peabody, Oliver, 1:27
- Peabody Picture Vocabulary Test (PPVT-III)**, 1:71, 2:748–749
- Pearson, Egon, 1:354, 2:750
- Pearson, E. S., 1:238
- Pearson, Karl**, 1:94, 133, 135, 354, 413, 2:530, 595, 617, 749–750, 783, 3:937, 938, 990
- Pearson Assessments, 2:609
- Pearson product-moment correlation coefficient**, 2:750–755
 - assumptions of, 2:751
 - construct validity, 1:178
 - descriptive statistics and, 2:751–753
 - effect size, 2:754–755
 - example of, 2:751–753
 - inferential statistics and, 2:753–754
 - interrater reliability, 2:485
 - meta-analysis, 2:596
 - nonparametric statistics, 2:687
 - predictive validity, 2:781–782
 - simple linear regression, 2:545
 - Spearman's rho versus, 1:189, 3:927, 933
 - validity coefficient, 3:1030–1031
- Pearson's biserial correlation, 1:96
- Pearson's chi-square. *See* **Chi-square test for independence**
- Pearson's r . *See* **Pearson product-moment correlation coefficient**
- Pedhazur, E. J., 1:179, 3:891
- Peirce, C. S., 1:421
- Penalty term, 1:15–16
- Percent agreement, 2:484
- Percentile and percentile rank**, 2:755–756
 - cautions about, 2:756
 - definitions and comparison, 2:755–756
 - normal curve, 2:694
- Perception
 - Aristotle's theory of, 2:707–708
 - Embedded Figures Test, 1:313–314
- Peres, S. Camille, 1:135
- Performance-based assessment**, 2:757–760
 - achievement tests and, 1:8
 - alternate assessment, 1:20–21
 - realism, 2:758–759
 - reliability, 2:759–760
 - uses, 2:758
 - validity, 2:759
 - varieties of, 2:757–758
- Performance IQ**, 2:757
- Peritz procedure**, 2:760–762
- Perry, C. L., 2:616
- Perry, J., 1:116
- Perry, M. A., 2:454
- Personal action construct units, 2:765
- Personality Assessment Inventory (PAI)**, 2:765–766, 769
- Personality disorders, 2:606
- Personality Inventory for Children, 2:769
- Personality Research Form (PRF)**, 2:766–767
- Personality tests**, 2:767–770
 - Basic Personality Inventory, 1:71–72
 - California Psychological Inventory, 1:117–118
 - Comrey Personality Scales, 1:175
 - Edwards Personal Preference Schedule, 1:299–300
 - Embedded Figures Test, 1:313–314
 - Holden Psychological Screening Inventory, 2:442
 - Jackson Personality Inventory–Revised, 2:498

- Millon Clinical Multiaxial Inventory-III, 2:606
 Minnesota Multiphasic Personality Inventory, 2:608–609
 Myers-Briggs Type Indicator, 2:675–676
 NEO Personality Inventory, 2:679–680
 nomothetic versus idiographic approaches, 2:684
 nonpathological, 2:769–770
 Personality Assessment Inventory, 2:765–766
 Personality Research Form, 2:766–767
 profile analysis, 2:788–789
 projective techniques, 2:768–769
 psychopathology assessment, 2:769
 purpose, 2:767
 self-report measures, 2:767–768
 Sixteen Personality Factor Questionnaire, 3:917
 Personal journals, 1:253
Personal projects analysis (PPA), 2:762–765
 assessment modules, 2:764
 criteria for, 2:762–764
 implications for research/practice, 2:765
 Person-item test of fit, 3:823
Perspectives on Psychological Science, 1:49
 Peskun, P., 2:571
 Pharmacology, 1:314
 Phase spectrum, 1:366
 Phillips, L. M., 2:439
 Philosophy
 ethics, 1:322–323
 Ockham's razor, 2:707–708
 Physiological attitude measures, 1:55–56
 Piaget, Jean, 3:950
 Picket fence effect, 1:368
Pie chart, 1:412, 2:770
Piers-Harris Children's Self-Concept Scale (PHCSCS), 2:770–771
 Pilkonis, P. A., 1:258
 Pilot study, value of, 3:856
 Piquero, A. R., 3:1043
 P.L. 94-142. *See* **Individuals with Disabilities Education Act**
 Placebo effect, 2:430
 Planck, Max, 1:117
 Planned comparisons, 2:646
 Platykurtic distribution, 2: 522
 Playfair, William, 1:412
 Plots
 exploratory data analysis, 1:328
 parallel coordinate, 2:727–729
 scree, 3:868–869
 stem-and-leaf, 1:380, 413, 414, 3:966–967
 sunflower, 3:977–978
 time, 3:1005
 Point biserial correlation, 1:94–96
 Point estimate, 3:818
 Point pattern data, 3:924–925
Poisson, Siméon Denis, 1:87, 2:771–772
 Poisson brackets, 2:772
 Poisson constant, 2:772
Poisson distribution, 2:772–775
 mean-variance equality, 2:772–773
 origins of, 2:772
 uses, 2:772
 Poisson integral, 2:772
 Poisson intensity function, 2:774
 Poisson ratio, 2:772
 Poisson regression, 3:830–831
 Polar-area diagrams, 1:413, 414
 Policy Delphi technique, 1:242
 Poll services, databases from, 3:872–873
 Polynomial regression, 1:212–213
 Population
 definition, 3:816
 Malthus and, 2:564
 parameter and, 2:732
 samples and, 2:457, 3:855, 941, 963
 substantive, 3:963
 Population-averaged models, 1:399–400
 Population median, 2:593
 Portability, of health insurance, 2:426–427
Portfolio assessment, 2:776–778
 alternate assessment, 1:20–21
 benefits, 2:776–777
 characteristics, 2:776
 development of portfolio, 2:777
 evaluation of portfolio, 2:777–778
 Positive definite matrices, 1:307
 Positive predictive value (PPV), 1:256
 Positive semi-definite matrices, 1:306–307, 3:849–850, 907–908
 Positivism, 1:418
 Postdictive design, 1:201
 Posterior Bayes factors, 1:75
Posterior distribution, 2:781
 Bayesian statistics, 1:80–81
 bivariate distributions, 1:98–99
 prior versus, 2:786–787
 Posterior distribution of θ , 1:80
Post hoc comparisons, 2:778–780
 honestly significant difference procedure, 2:778–780

- planned versus, 2:646
- Scheffé procedure, 2:778–780
- Posttest score, 1:258
- Potency dimension of connotative meaning, 3:878, 881
- Power. *See* Statistical power
- Power analysis, 3:1022–1023
- Power functions, 1:213–214, 3:863
- Power spectrum of the signal, 1:366
- Power test, 2:517
- Practical modeling, 1:232–233
- Practical power, 1:155
- Pragmatism, 1:418, 420
- Praxis, 1:298–299
- Precision, 1:97
- Predicted variable, 2:624
- Prediction
 - ability tests, 1:4
 - aptitude tests, 1:39–40, 42
 - bivariate distributions, 1:97–98
 - causal analysis, 1:123
 - differential, 2:492
 - mixed-effects models, 2:615
 - predictive validity, 1:201
 - regression, 3:831
 - simple linear regression, 2:545
 - See also* Forecasting
- Prediction bias, 2:492
- Predictive equations, 2:545
- Predictive validity**, 1:200–201, 2:781–782, 3:1029, 1033
- Predictors of change, 1:259–260
- Predictor space, 1:265
- Predictor variable
 - criterion validity and, 1:200, 202
 - fixed versus random, 2:612
 - homogeneity of variance, 2:442
 - nonexperimental research, 2:624
 - simple linear regression, 2:545
 - suppressor variable and, 3:980
- Preexisting conditions, health insurance and, 2:426–427
- Pre-post design, 3:836
- Pre-posttest design, 3:838
- Preschool Language Assessment Instrument (PLAI-2)**, 2:782–783
- Press, S. James, 2:501
- Pretests, covariates and, 1:31
- Pretest score, 1:258
- Primacy effects, 3:810
- Primary data, 1:220
- Primary Mental Abilities, 2:477
- Primary sampling units, 1:150
- Principal axis factors, 1:334
- Principal component analysis (PCA)**, 2:783–786
 - curse of dimensionality, 1:210–211
 - dimension reduction, 1:265
 - example of, 2:784–785
 - exploratory factor analysis, 1:334
 - factor scores, 1:348
 - formulation of, 2:783–784
 - generalizations of, 2:785–786
 - goal of, 2:783
 - kernel, 2:786
 - local, 2:785–786
 - metric multidimensional scaling, 2:598
 - missing data method, 2:610–611
 - multiple factor analysis, 2:657–663
 - origins of, 2:783
 - orthogonal predictors, 2:717
 - partial least square regression, 2:740
 - scree plots, 3:868–869
 - singular value decomposition and, 3:907, 911–912
 - STATIS, 3:955–962
- Principal component regression, 2:740–741
- Principal components, 2:784
- Principal curve, 2:785
- Principal Hessian directions (pHd), 1:266
- Prior distribution**, 2:786–787
 - Bayesian statistics, 1:80–81
 - bivariate distributions, 1:98
 - informative versus noninformative, 1:80
- Privacy
 - health insurance and, 2:427–428
 - protecting respondent, 3:1000–1001, 1012–1013
 - student disabilities and, 2:455
- Probability
 - Bayes' theorem, 1:79
 - Bernoulli and, 1:86–87
 - binomial distribution/binomial and sign tests, 1:87–90
 - causality, 1:124
 - chance and, 1:129–130
 - conditional, 1:175–176
 - exact, 1:89, 137
 - Fisher exact probability test, 1:354–358
 - Poisson and, 2:772
 - Poisson distribution, 2:772–775
 - sample weights and, 3:981
 - tree diagrams, 3:1012–1013
 - Type II error as, 3:1021–1022
- Probability density function (PDF), 2:690–692, 3:982
- Probability models, 2:446–447
- Probability sampling**, 2:787

- cluster sampling, 1:150
- definition, 1:187, 2:689
- nonprobability sampling versus, 3:812–813
- Probability scoring, 2:460
- Probability theory
 - chance, 1:129
 - inferential statistics, 2:457
- Probes, 1:206
- Probit regression, 3:830
- Problem solving
 - curriculum-based measurement for, 1:207–208
 - Future Problem Solving Program, 3:1010
 - grounded theory and, 1:418
 - Response to Intervention and, 3:840
 - Six Sigma, 3:914
- Procedural justice, in research, 1:324
- Processing speed, 1:3, 406, 2:478
- Procrustes matching by congruence coefficients, 3:955
- Product moment, 1:97
- Product-moment correlation, 1:360
- Product terms, 1:48
- PRO-ED, 2:782
- Profile, 1:281
- Profile analysis, 2:788–789**
- Profile of Mood States, 2:605
- Progressive Matrices Test, 1:204–205
- Projection pursuit, 1:265
- Projective Hand Test, 2:789**
- Projective hypothesis, 2:511
- Projective testing, 2:789–793**
 - common types, 2:769, 790–792
 - ethics of, 2:792
 - figure drawing tests, 2:791
 - Hand Test, 2:789
 - origins of, 2:768
 - personality tests, 2:768–769
 - principles of, 2:768, 790
 - resurgence of, 2:792–793
 - Roberts Apperception Test for Children, 3:847
 - Rorschach Inkblot Test, 3:848–849
 - self-report measures versus, 2:769
 - sentence completion tests, 2:791–792
 - test bias, 2:769
 - Thematic Apperception Test, 3:1000
 - validity, 2:769
 - value of, 2:789–790
- Promoter sites, 1:91
- Proof, hypothesis testing and, 2:448
- Propensity scores, 2:705, 793–795, 3:807**
- Proportional allocation, 3:969–970
- Proportional hazards model. *See* Cox proportional hazard model
- Proportion of common variance, 1:160
- Proposal phase, 3:897
- ProQolid Quality of Life Instruments Database, 3:1050–1051
- ProQuest Digital Dissertations, 1:25
- ProQuest Information and Learning database, 1:25
- Proschan, F., 2:535
- Prospective studies, 3:832
- Protected health information (PHI), 2:427–429
- P set, 3:800
- Pseudo-guessing parameter, 2:493
- Pseudo random number generators, 2:627, 3:815
- Psychoanalysis, projective techniques and, 2:768, 792
- Psychological Abstracts, 2:795–796**
- Psychological Assessment Resources, 1:138, 2:679, 770
- Psychological Bulletin*, 2:500
- Psychological Corporation, 1:4, 81, 82, 260, 299, 406, 2:607, 3:988, 1049
- Psychological Review*, 2:500
- Psychological Science*, 1:49
- Psychological Science in the Public Interest*, 1:49
- Psychology, evidence-based practice in, 1:326
- Psychology Assessment Resources, 2:765
- Psychometrics, 2:796–797**
- Psychometrics, Jackson and, 2:497
- Psychopathology assessment, 2:442, 769
- PsycINFO, 2:797–798**
- Publication bias, 1:353, 2:596, 597
- Publication Manual* (American Psychological Association), 1:178, 3:891
- Publications of the American Statistical Association*, 1:27, 2:500
- Purification subtest, 2:491
- p* values, 2:534–535
- Q methodology, 3:799–802**
 - concourse, 3:799
 - factor interpretation, 3:801–802
 - P set, 3:800
 - Q samples, 3:799
 - Q sorting, 3:800–801
 - R versus, 3:801
 - single cases, 3:802
 - statistical procedures, 3:801
- Q-Q plot, 3:802–804**
 - example of, 3:803
 - uses, 3:802–803
- Q samples, 3:799

- Q-Sort technique, 1:245, 3:800–801
- Quade test, 1:381
- Quadratic entropy, 1:278–279
- Qualitative data, 1:220
- Qualitative research
 - audit trails, 1: 61
 - descriptive research and, 1:250–251
 - grounded theory, 1:418–421
 - secondary data analysis, 3:870–872
 - text analysis, 3:999
- Quality-adjusted life years (QALYs), 3:804
- Quality function deployment, 3:915
- Quality management, 1:248–249, 3:912–917
- Quality of Well-Being Scale (QWB), 3:804–805**
- Quality of Well-Being Scale–Self-Administered (QWB–SA), 3:805
- Quantification method, 2:651
- Quantile-quantile plot. *See* **Q-Q plot**
- Quantitative data, 1:220
- Quantitative knowledge (Gq), 1:3, 404
- Quantitative research
 - descriptive research, 1:250–251
 - secondary data analysis, 3:870–872
 - text analysis, 3:999
- Quarterly Publications of the American Statistical Association*, 2:500
- Quartiles
 - box plot, 1:112
 - normal curve, 2:694
 - percentiles and, 2:756
- Quasi-experimental method, 3:805–808**
 - description of, 3:805
 - examples of, 3:807
 - modifications of, 3:807
 - statistical adjustments, 3:807–808
 - tests of mediating effects, 3:995–997
 - types, 3:806
 - validity, 3:805–806*See also* **Observational studies**
- Quasi-likelihood, 1:398
- Quasi-likelihood information criterion, 1:400
- Questionnaires, 3:808–811**
 - administration methods, 3:808–809
 - construction of, 3:811
 - descriptive research, 1:251
 - interviewer-administered, 3:809
 - labeling scale points, 3:810
 - midpoints, 3:810–811
 - nonresponse options, 3:811
 - number of scale points, 3:810
 - open- versus closed-ended questions, 3:809
 - order effects, 3:810, 811
 - question construction, 3:809–811
 - rating versus rank ordering, 3:810
 - self-administered, 3:808–809
 - thematically organized versus randomized, 3:811
 - wording of questions, 3:809–810
- Questions
 - double-barreled, 3:809
 - open- versus closed-ended, 1:315, 3:809
 - wording of, 3:809–810
- Quételet, Adolph, 2:441
- Quiz, definition of, 1:8
- Quota sampling, 2:689, 3:812–814**
- Race
 - anthropometry and, 1:36
 - System of Multicultural Pluralistic Assessment, 3:988*See also* Ethnicity
- Racial profiling, 1:257
- Radial basis function networks, 1:234
- Raghunathan, Trivellore, 2:664
- Rajaratnam, Nageswari, 1:157
- Ramsey, P. H., 1:154
- Ramsey, P. P., 1:154
- RAND Corporation, 1:240
- Random assignment, random sampling versus, 3:819
- Random-coefficient regression modeling. *See* **Hierarchical linear modeling**
- Random-effects models, 1:272–275, 399
 - See also* **Hierarchical linear modeling**
- Random error, 1:141, 2:585, 730, 3:933
- Random generation process, 2:446–447
- Randomization, random sampling versus, 3:819
- Randomization test, for clusters, 2:504–505
- Randomized block fractional factorial design. *See* **Fractional randomized block design**
- Randomized block *F* test, 1:151
- Randomized complete block designs
 - Cochran *Q* test, 1:151, 152
 - Friedman test, 1:380
 - Page’s *L* test, 2:719–722
- Randomized experiments. *See* Experiments
- Randomized-response techniques, 3:1001
- Randomness
 - gambler’s fallacy and, 1:384
 - stochastic processes, 2:568, 3:901
- Random number generation
 - Monte Carlo methods, 1:169
 - pseudo, 2:627, 3:815

- testing, 3:816
- true, 3:815
- Random numbers, 3:815–816**
- Random sampling, 3:816–819**
 - estimation of parameters, 3:817–818
 - parametric hypothesis testing, 3:818
 - randomization versus, 3:819
 - representativeness, 3:816–817
 - simple, 3:816
 - with replacement/without replacement, 3:816–817
- Random sampling distribution of differences, 3:964
- Range, 3:820**
- Rank, of matrix, 1:308
- Rank-deficient matrix, 1:308
- Rank ordering, 3:810
- Rank tests, 2:686
- Rao, C. R., 1:278, 311
- Rasch, George, 3:821, 864
- Rasch measurement model, 3:820–825**
 - additivity concept in measurement, 2:583–584
 - computerized adaptive testing, 3:824–825
 - data analysis with RUMM computer program, 3:823–824
 - generalized linear mixed models, 2:615
 - item banks, 3:824–825
 - item response theory, 2:494
 - many-facets model (MFMR), 3:824
 - measurement in, 3:821
 - partial credit model, 3:822–823
 - raw score conversion to linear scale, 3:820–821
 - simple logistic model, 3:821
 - uses, 3:820
- Rasch Unidimensional Measurement Models (RUMM), 3:822–824
- Raters, 1:7, 3:1002–1005
 - See also Interrater reliability*
- Rating scales, 1:21, 3:810
- Ratio estimators, 1:248
- Ratio level of measurement, 3:825–826**
 - absolute zero in, 3:825
 - definition, 2:582–583
 - interval versus, 2:486–487
- Raven's Progressive Matrices, 2:478, 3:826–827**
- Raw score, 1:9, 3:820–821, 952–954
- Reaction time/speed, 1:3
- Reaction-time theory, 2:478
- Reactive Delphi technique, 1:243
- Reading/writing ability (Grw), 1:404
- Realism, 1:419, 2:758–759, 3:1035
- Real-time Delphi technique, 1:242
- Reasonable accommodations, 3:876–877
- Receiving operator characteristic curve (ROC), 1:255
- Recency effects, 3:810
- Reciprocal causality, 1:124
- Reciprocal loops, 3:974
- Reciprocal suppression, 3:980
- Reconstructed proportion, 3:910
- Record linkage, 2:527, 3:827–828**
- Recurrent networks, 1:47
- Reed Elsevier, 3:950
- Reference data. *See* Training data
- Reference groups, 2:490, 3:952
- Refined factor scores, 1:346–347
- Regression
 - Galton and, 1:383
 - least squares method, 2:530–532
 - path analysis and, 2:746
 - Pearson and, 2:750
 - See also Curvilinear regression*; Multiple regression analysis
- Regression analysis, 3:829–832**
 - categorical outcomes, 3:830
 - censored outcomes, 3:831
 - control, 3:831–832
 - count outcomes, 3:830–831
 - explanation, 3:831
 - linear, 2:544–550
 - linear algebra algorithms, 1:168
 - logistic, 2:550–554
 - nonparametric methods, 2:688
 - orthogonal predictors, 2:717–718
 - prediction, 3:831
 - semiparametric, 3:986–987
 - significance of, 2:545
 - types, 3:830–831
 - uses, 3:831–832
- Regression coefficient, 2:549
- Regression constant, 2:549
- Regression-discontinuity quasi-experiments, 1:29, 3:806
- Regression line, 2:545, 3:829
- Regression slopes, 1:32
- Regression-to-the-mean, 1:42–43
- Regularity theory, 3:937
- Rehabilitation Act of 1973, Section 504, 1:27–28, 3:876–877
- Rejection sampling, 2:569
- Relative risk (RR), 3:832–833**
- Relaxation time, 3:898
- Reliability
 - achievement tests, 1:8

- alternate form, **1**:155
- aspects of, **3**:834–835
- benefits of, **3**:993
- Carroll Depression Scale, **1**:121
- classical test theory, **1**:140
- coefficient alpha, **1**:155–157
- correction for attenuation, **1**:51–52
- Differential Aptitude Tests (DAT), **1**:4
- error and, **2**:730, **3**:933–934
- estimation of, **1**:155, **3**:993
- ethics of testing selection and, **1**:319
- increasing, strategies for, **3**:945
- issues regarding, **2**:731, **3**:935, 994
- KR-20 and KR-21, **2**:516–519
- parallel forms, **2**:730–731, **3**:993
- performance-based assessment, **2**:759–760
- portability of, **2**:516
- reliability coefficient versus standard error of measurement, **3**:944
- software packages, **2**:519
- split half, **2**:516–518, **3**:835, 934–935
- test length, **3**:935
- test-retest, **1**:155, 157, **3**:834, 944, 992–995
- true score and, **3**:1015
- validity and, **1**:178, **3**:834, 1031
- See also* **Interrater reliability**; **Reliability theory**
- Reliability coefficient, **3**:944
- Reliability generalization, **2**:516
- Reliability theory, 3:834–835, 1015**
- See also* Reliability
- Religion, Ockham's razor and, **2**:707–708
- Relles, D. A., **1**:59–60
- Rencher, A. C., **3**:804
- Rennie, David, **1**:419, 421
- Repeated measures analysis of variance, 1:33, 3:835–838**
- between-subjects factors designs, **2**:559
- example of, **3**:837–838
- sphericity and, **1**:35
- Repeated measures data. *See* **Longitudinal/repeated measures data**
- Repeated measures design. *See* **Longitudinal/repeated measures data**
- Replacement, sampling with/without, **3**:816–817
- Representativeness, of samples, **2**:457, **3**:817, 941
- Reproducibility coefficient, **1**:423
- Reproducibility studies, **3**:916
- Research
 - application and, **1**:72–73
 - applied, **1**:39
 - audit trails, **1**: 60–61
 - basic, **1**:39, 72–73
 - Delphi technique, **1**:241, 244
 - descriptive, **1**:250–253
 - ethics, **1**:321–325, **2**:465–468
 - life course of, **2**:698
 - meta-analysis of, **2**:595–598
 - responsibilities in, **1**:323
 - secondary data analysis, **3**:870–871, 875
 - subjects of, **1**:222–224
 - theory and, **1**:72–73
- Researchers
 - data collection, **1**:222–225
 - ethics, **1**:323–325
- Research problem, **1**:222
- Research Psychologists Press, **2**:499
- Research subjects, **1**:222–224
- Residuals, 1: 62, 134, 2:541, 737–738, 3:838–840**
- definition, **3**:838
- illustration of, **3**:839
- types, **3**:839–840
- uses, **3**:838–839
- Respect, in research, **1**:324–325
- Response bias, **2**:497
- Response items, selected versus constructed, **1**:7
- Response process validity, **3**:949
- Response rates, **1**:251
- Response set, **3**:843
- Response to Intervention (RTI), 3:840–843**
- advantages, **3**:842
- background, **3**:840–841
- characteristics, **3**:841
- controversies over, **3**:842–843
- curriculum-based measurement, **1**:208, **3**:841
- Gresham and, **1**:417
- pedagogical value of, **3**:841–842
- three-tier model, **3**:841
- Restricted maximum likelihood (REML), **2**:614–615
- Restriction of range, **3**:1031
- Retrospective studies, **3**:833
- Retzius, Anders, **1**:36
- Reverse coded items, **3**:843
- Reverse scaling, 3:843–845**
- cautions about, **3**:845
- coding and scoring, **3**:844–845
- definition, **3**:843
- purpose, **3**:843
- semantic differential method, **3**:879
- SPSS recoding, **3**:844–845
- Revised Children's Manifest Anxiety Scale, **3**:846

- Revisions, of tests, 1:319
- Reynolds, Cecil R., 3:846**
- Reynolds, S. K., 1:258
- Reynolds Intellectual Assessment Scales, 3:846
- r* family, 2:596
- Ricardo, David, 2:564
- Richardson, K., 2:479
- Richardson, M. W., 1:155
- Richardson extrapolation, 3:895
- Rider, R. A., 3:837
- Right censoring, 1:125, 3:982–983
- Ripley's K function, 3:925
- Risk-benefit ratio. *See* Cost-benefit considerations
- Risk difference, 1:57
- Riverside Publishing Company, 1:85, 162, 2:487, 488
- R methodology, 3:801
- Robbins-Monro process, 1:171
- Robert, P., 3:849
- Roberts Apperception Test for Children (RATC), 3:847**
- Robeson, Paul, 3:1051
- Robust variance estimator, 1:399
- Rod and Frame Test, 1:313
- Roentgen-cephalometry, 1:35
- Romero, D. W., 1:24
- Root mean square error of approximation, 3:975
- Roper Center for Public Opinion Research, 3:872
- Rorschach Inkblot Test, 2:769, 790, 792–793, 3:848–849**
- Rosenbluth, A., 3:896
- Rosenbluth, M., 3:896
- Rosenthal, R., 1:300, 301, 303, 353, 2:595, 596, 597
- Rosopa, P., 3:996–998
- Ross, Frank A., 2:500
- Rotation, 1:335
- Rotter, Julian B., 2:479, 791
- Royal Astronomical Society, 1:69
- Royal Society, 1:69
- Royal Society of Edinburgh, 1:69
- Royal Statistical Society, 1:69, 413
- Roy's greatest characteristic root, 2:670–671
- Royston, J. P., 3:884
- r*-scan, 3:865–866
- R statistical software system, 1:170, 3:979
- Rubin, D., 1:310
- Rubin, D. B., 1:261, 263, 2:595
- Rubin, H., 2:471
- Rubrics
- achievement tests, 1:8
 - performance-based assessment, 2:759–760
- Rule induction, 1:234
- Russell, Bertrand, 1:123
- Russian papyrus, 3:896
- R_v and congruence coefficients, 3:849–852**
- example of, 3:851–852
 - notations and computational formulas, 3:849–850
 - sampling distributions, 3:850–852
- Rychlec, M., 2:430
- Sabatier, R., 3:851
- Sacramento City Unified School Dist. Bd. of Educ. v. Rachel H.*, 2:456
- Sage, 3:920
- Salih, Fathi A., 1:157
- Salsburg, D., 1:328–329
- Sample, 3:855–856**
- definition, 3:816
 - population and, 2:457, 3:855, 941, 963
- Sample covariances, 1:194
- Sample median, 1:316–318, 2:593
- Sample size, 3:856–860**
- acceptance sampling, 1:7
 - allocation, 3:859
 - analysis of variance, 3:858–859
 - effect size, 3:859–860
 - error and, 3:994
 - importance of, 3:856
 - power requirements, 3:857–858
 - process of determining, 3:856
 - significance level, 3:890–891
 - software packages, 3:858
 - specified confidence interval width, 3:856–857
 - structural equation modeling, 3:974
- Sampling
- acceptance, 1:6–7, 2:444–445
 - adaptive, 1:12
 - chance and, 1:131
 - cluster, 1:150–151
 - convenience, 1:186–188, 2:689
 - definition, 3:816
 - nonprobability, 2:689–690
 - population and, 2:457
 - probability, 2:787
 - quota, 2:689, 3:812–814
 - random, 3:816–819
 - rejection, 2:569
 - stratified, 1:12
 - theoretical, 1:419
 - with replacement/without replacement, 3:816–817, 855–856
- Sampling distribution of a statistic, 2:447, 3:860–861**

Sampling error, 3:861–862

Sampling scheme, 3:855

Sampling weights, 3:981

Sándor, Aniko, 1:135

Sandwich variance estimator, 1:399

SAS Enterprise Miner, 1:233

SAS/IML, 1:170

SAS JMP statistical software package, 1:137–138

SAS statistical software package

data mining, 1:233

multiple imputation for missing data, 2:664

polynomial regression, 1:213

Satisficing bias, 3:843–844

Satterthwaite's approximation, 1:84–85

Saunders, D. R., 1:48

Savickas, Mark L., 1:120

Sawilowsky, Shlomo S., 1:179, 317, 2:500–501,
629–630

Scale parameter, 2:690

Scale points, 3:810

Scales, definition of, 3:843

Scale scores, 1:347

Scaling, 3:862–865

category, 3:864

definition, 3:862

difference-based, 3:864

discriminability, 3:862–863

item response theory, 3:864–865

magnitude, 3:863–864

origins of, 3:862

Scalogram analysis, 1:421–424, 2:651

See also **Guttman scaling****Scan statistic, 3:865–867**

assumptions of, 3:867

computations, 3:867

distribution function of, 3:865

hypothesis testing, 3:867

molecular biology application, 3:866–867

r-scan, 3:865–866 *w*-scan, 3:866**Scattergram, 3:868**

simple linear regression, 2:545

sunflower plot, 3:977

Schafer, Joe, 2:664

Schapire, R., 1:270

Scheffé procedure, 2:647, 778–780

Schmelkin, L. P., 1:179, 3:891

Schmidt, F. L., 1:48, 2:595

Scholastic Aptitude Test (SAT), 1:41, 42, 298, 411, 2:529

Scholastic Assessment Test, 1:42

Scholastic Testing Service, 3:1010, 1011

School Readiness Composite (SRC), 1:113

Schrödinger, Erwin, 2:627

Schrödinger equation, 2:627

Schroër, G., 2:515

Schuler, R., 1:244

Schur scalar product, 3:850

Schwarz criterion, 1:75

Science, Ockham's razor and modern, 2:708–709

Score equating/linking, 2:495–496

Score matrix, 2:741

Scree plot, 1:333, 3:868–869

Searles, John S., 2:595

Seasonality, 3:1005

Seating sweeps method, 1:252

Sechrist, S. R., 1:244

Secondary data, 1:220

Secondary data analysis, 3:869–876

advantages, 3:874

challenges, 3:874–875

databases, 3:872–874

efficiency of, 3:870, 874

ethical issues, 3:874

meta-analysis versus, 3:869–870

practical basis for, 3:870

process, 3:871–872

protection of research pool, 3:870, 874

qualitative uses, 3:870–872

quantitative uses, 3:870–872

research applications, 3:875

research uses, 3:870–871

special research needs and, 3:870

strengthened confidence from, 3:870

unobtrusive measures, 3:874

Secondary sampling units, 1:150

**Section 504 of the Rehabilitation Act of 1973,
1:27–28, 3:876–877**

See, W. Y., 1:65

Selected response items, 1:7

Selection bias, probability versus nonprobability
 sampling, 2:689, 3:812–813

Selection bias modeling, 3:807

Selective coding, 1:419

Self-administered questionnaires, 3:808–809

Self-concept measures, 2:770–771

Self-Determination Knowledge Scale (SDKS), 2:518

Self-organizing maps, 1:234

Self-report, 3:877–878

construction of, 2:767–768, 3:877–878

personality tests, 2:767–768

- projective techniques versus, 2:769
- Quality of Well-Being Scale, 3:805
- test bias, 2:768, 3:877–878
- Self-similarity, 1:370–373
- Semantic differential, 3:878–881**
 - attitude tests, 1:54
 - definition, 3:878
 - development of, 3:878
 - Likert scaling, 2:539
 - scale construction, 3:878–880
 - strengths, 3:880
 - weaknesses, 3:880–881
- Semantic differential scale, 3:881–882**
- Semi-interquartile range, 3:882–883**
- Semiparametric regression analysis, 3:986–987
- Semi-partial regression coefficient. *See* **Part and partial correlations**
- Sen, Pranab K., 2:501
- Sensitive questions, technique for asking, 3:1000–1001
- Sensitivity, 1:201–202, 255–256, 263, 2:706
- Sentence completion projective tests, 2:791–792
- Serlin, Ronald C., 2:501
- Shaffer, Juliet P., 2:501, 681
- Shanmugam, R., 2:773–775
- Shapiro, S. S., 3:884
- Shapiro-Francia W' test, 3:885
- Shapiro-Wilk test for normality, 3:883–885**
 - drawbacks, 3:884
 - example of, 3:884–885
 - power, 3:885
 - software packages, 3:883–885
- Shattuck, Lemuel, 1:27
- Shaw, George Bernard, 3:989
- Shelley, G. A., 1:127
- Shepard, Roger, 3:864
- Shermer, M. B., 1:218
- Sherry, A., 1:15
- Shewhart, Walter, 1:248
- Shields, C. M., 1:225
- Shih, W. Joe, 1:95
- Shlaes, J. L., 1:56
- Short, Thomas, 2:502
- Short-term memory, 1:3, 405, 3:951
- Shrout, P. E., 3:996, 998
- Shrunken value, 1:158, 161, 2:649
- Shuford, Emir, 1:114
- Shulkin, B., 1:317
- Shuo, C., 1:65
- Šidák, Zbynek, 1:294
- Šidák equation, 1:105–106
- Sierpinski Gasket, 1:370, 372, 374
- Sigma Assessment Systems, 1:71, 2:497, 498, 640, 766
- Sigmoid firing rule, 1:47
- Signal-contingent schedules, 1:297
- Signal detection theory (SDT), 3:886–889**
 - application of, 3:889
 - definition, 3:886
 - example of, 3:886–887
 - goal of, 3:886
 - model, 3:887–889
 - threshold, 3:888
 - uses, 3:886
- Signed rank tests, 2:686, 723, 3:1051–1053
- Significance. *See* **Statistical significance**
- Significance level, 3:889–891**, 966, 1019
- Significance testing
 - equivalence testing versus, 1:314–315
 - hypothesis testing, 2:446–448
 - implications, 3:891
 - inferential statistics, 2:458
 - Kendall rank correlation, 2:509–510
 - null hypothesis, 2:695–698
 - significance level, 3:889–891
 - uses, 3:890
- Sign tests, 1:88–89, 380, 2:686
- Silverberg, K. E., 3:875
- Simon, Herbert, 1:123, 3:937, 939
- Simon, Théodore, 1:1, 39, 2:474, 3:950
- Simple linear regression, 2:545–546
 - assumptions of, 2:546
 - error scores, 2:545–546
 - predictive equations, 2:545
- Simple logistic model (SLM) of Rasch, 3:821–822
- Simple logistic regression model, 2:551–553
- Simple main effect, 3:892**
- Simple random sampling, 2:787, 3:816
 - cluster sampling versus, 1:150–151
 - stratified random sampling versus, 3:968
- Simple structure, 1:335
- Simpson, Thomas, 3:896
- Simpson's paradox, 2:765, 3:892–894**
- Simpson's rule, 3:895–896**
- Simulated annealing (SA), 2:570–571, 3:896–900**
- Simulation experiments, 3:901–903**
- Simulation programming language, 3:903
- Single-case studies. *See* **Single-subject designs**
- Single-item direct attitude measures, 1:54
- Single-subject designs, 3:903–907**
 - baseline phase, 3:904–905
 - between-subjects versus, 3:904

- characteristics, 3:904
- multiple-baseline designs, 3:906–907
- Q methodology, 3:802
- quasi-experiments, 3:806
- uses, 3:904
- withdrawal designs, 3:905–906
- Singular and generalized singular value decomposition, 3:907–912**
 - definitions and notations, 3:907–909
 - image compression, 3:911–912
 - mathematical properties, 3:910–911
- Singular matrix, 1:308
- Singular value decomposition. *See* **Singular and generalized singular value decomposition**
- Singular values, of matrix, 1:309
- Six Sigma, 3:912–917**
 - assumptions of, 3:914
 - meanings of, 3:912–913
 - methodology of, 3:914–917
 - metric of, 3:913–914
- Sixteen Personality Factor Questionnaire (16PF), 2:770, 3:917–918**
- Sixteenth Mental Measurement Yearbook*, 1:10
- Skewness, 2:674, 3:885, 918**
- Skills, achievement and, 1:8
- Sliced average variance estimation (SAVE), 1:266–267
- Sliced inverse regression (SIR), 1:266–267
- Slope, 2:549
- Smith, M. A., 1:329–331
- Smith, Mary L., 2:595
- Smoothing, 2:633–637, 3:918–919**
- Smythe, W. E., 1:128, 2:585
- Snow, John, 1:412
- Social Behavior Assessment, 3:921
- Social Climate Scales, 3:919–920**
- Social constructionism, 1:418
- Social desirability effects, 2:430, 3:809
- Social indicators, personal projects analysis and, 2:763
- Social Learning Theory, 2:479
- Social Policy Report*, 3:922
- Social Skills Rating System (SSRS), 1:417, 3:920–921**
- Society for Multivariate Experimental Psychology, 1:174
- Society for Research in Child Development (SRCD), 3:921–922**
- Sociological Abstracts, 3:923**
- Soft margins, 3:979
- Softmax function, 2:622
- Software
 - Excel spreadsheet functions, 1:326–328
 - exploratory data analysis, 1:329
 - F* test, 3:837
 - generalized estimating equations, 1:400
 - hierarchical linear modeling, 2:435
 - Information Referenced Testing, 2:460
 - interactive, 1:414–416
 - likelihood ratio test, 2:535
 - mixed-effects models, 2:615
 - Monte Carlo option, 2:629
 - multiple imputation for missing data, 2:664
 - nonparametric methods, 2:688
 - normal curve, 2:694–695
 - open source movement, 1:170
 - partial least square regression, 2:744
 - path analysis, 2:748
 - reliability estimates, 2:519
 - sample size, 3:858
 - Shapiro-Wilk test for normality, 3:883–885
 - spreadsheet functions, 3:935–937
 - statistical graphics, 1:414–416
 - structural equation modeling, 3:974, 975
 - support vector machines, 3:979
 - text analysis, 3:999
 - time series analysis, 3:1009
 - t* test for two population means, 3:991–992
 - Tukey-Kramer procedure, 3:1018
 - variance, 3:1042–1043
 - See also specific programs*
- Sone scale of loudness, 3:863
- Soper, H. E., 1:96
- Sorting distance, 1:280
- Southeastern Community College v. Davis*, 3:877
- Space-alternating generalized EM algorithm, 1:313
- Spargo, P. E., 1:199
- Sparrow, Sara, 3:1045
- Spatial autocorrelation, 3:925–926
- Spatial data, 3:924–926
- Spatial Learning Ability Test (SLAT), 3:923–924**
- Spatial statistics, 3:924–927**
 - geostatistical data, 3:925–926
 - lattice data, 3:926
 - point pattern data, 3:924–925
- Spearman, Charles, 1:2, 50, 50, 142, 341, 404, 2:477, 3:927, 951, 1014
- Spearman-Brown prophecy formula, 1:155, 2:519, 3:935, 994
- Spearman's *g*. *See* General intelligence (*g*)
- Spearman's rho, 3:927–933**
 - computation, 3:927–928

- correlation coefficients, 1:189, 3:927
- descriptive statistics and, 3:928–929
- inferential statistics and, 3:929–932
- nonparametric statistics, 2:687
- Pearson's *r* versus, 3:927, 933
- Specification limits, 3:913
- Specificity, 1:201–202, 255–256
- Specific learning disabilities (SLDs), 3:840–843
- Specific objectivity, 2:583–584
- Spectral analysis, 1:365
- Spectral decomposition, 1:309
- Spectral leakage, 1:369
- Spectrum, 1:304
- Speech, data compression of, 1:228–231
- Speed test, 2:517
- Sphericity, 1:35, 2:560, 716, 3:836, 851
- Spiessens, B., 1:263
- Spineplot, 2:631
- Splice site, 1:91
- Split half reliability**, 2:516–518, 3:835, 933–935
- S Plus statistical software package, 1:170, 2:664
- Spreadsheet functions**, 1:326–328, 3:935–937
- SPSS Clementine, 1:233
- SPSS statistical software package
 - chi-square goodness-of-fit, 1:135
 - classification trees, 1:143
 - coefficient alpha, 1:156–157
 - Cohen's kappa, 1:165
 - data mining, 1:233
 - difference scores, 1:259
 - inferential statistics, 2:458
 - interrater reliability, 2:484–485
 - logistic regression analysis, 2:552–553
 - loglinear analysis, 2:555–558
 - longitudinal/repeated measures data, 2:560
 - Mann-Whitney U test, 2:567
 - Mean command, 1:44
 - measures of central tendency, 2:589–590
 - median, 2:592
 - median test, 2:594
 - missing data method, 2:611–612
 - paired samples *t* test, 2:724–725
 - regression analysis, 3:830
 - reliability estimates, 2:519
 - reverse scaling, 3:844–845
 - standard error of the mean, 3:943
 - Wilcoxon Signed Ranks Test, 3:1053
- Spurious correlation**, 3:937–939
 - causation and, 3:937, 939
 - correlation types and comparison to, 3:938
 - definition, 3:938
 - generative causation, 3:938–939
 - senses of, 3:937
- Squared error loss, 1:237
- Squared multiple correlation coefficient. *See* **Multiple correlation coefficient**
- Stability, of data, 3:904
- Stability-equivalence estimate of reliability, 3:994
- Staelin, Richard, 2:608
- Standard deviation**, 3:940–941
 - effect size, 3:859
 - normal curve, 2:692
 - Pearson and, 2:750
 - standard error of the mean and, 3:941
 - variance and, 3:1042
 - See also* **Deviation score**
- Standard error. *See* **Standard error of the mean**
- Standard error of differences, 3:964
- Standard error of measurement (SEM)**, 3:944–945
 - confidence intervals, 3:945
 - definition, 3:944
 - estimation of, 2:730, 3:934, 993
 - sampling error, 3:861–862
 - test construction, 1:199
 - uses, 3:944–945
- Standard error of the mean**, 3:941–943
 - assumptions of, 3:942
 - example of, 3:942–943
 - SPSS calculation of, 3:943
- Standardized achievement tests
 - assessment versus high-stakes, 1:10
 - characteristics, 1:10
 - criterion-referenced tests, 1:195
 - purposes, 1:10
 - specific versus general, 1:10
- Standardized mean difference, 1:301
- Standardized regression coefficients, 2:549
- Standard scores**, 1:360, 3:945–946
 - See also* **T scores; z scores**
- Standard settings, 1:21
- Standards for Educational and Psychological Testing**, 1:8, 10, 183, 338, 2:490, 3:946–949, 950, 1029
 - organization of, 3:947–948
 - purpose, 3:947
 - second versus third editions, 3:948–949
 - validity concept in, 3:949
- Stanford Achievement Test**, 1:10, 3:949–950
- Stanford-Binet Intelligence Scales**, 1:1, 39, 71, 2:475, 476–478, 3:950–951
- Stanine**, 3:951–955

- advantages/disadvantages, 3:954
 - cautions about, 3:955
 - determination of, 3:951–952
 - labeling, 3:952
 - meaning and origin of word, 3:952
 - norm-referenced tests, 3:953
 - raw scores and, 3:952–954
 - uses, 3:954
- Stanley, Julian C., 3:805
- State-Trait Anger Expression Inventory (STAXI), 2:769
- State-Trait Anxiety Inventory (STAI), 1:121, 2:605, 769
- Static systems, 3:901
- STATIS**, 1:284, 3:955–962
- analyzing compromise, 3:959–960
 - computing compromise matrix, 3:957–959
 - example of, 3:956
 - goal of, 3:955
 - notations, 3:956–957
 - original variable integration, 3:962
 - projecting into compromise space, 3:960–962
 - R_v and congruence coefficients, 3:849–852
 - uses, 3:955
- Statistic, parameter versus, 2:732, 3:963
- Statistical analysis, data mining versus, 1:232–233
- Statistical computing, 1:168–169
- evaluation of special functions, 1:169
 - numerical linear algebra, 1:168
 - numerical optimization, 1:168–169
 - random number generation, 1:169
- Statistical conclusion validity, 3:806
- Statistical Decision Functions* (Wald), 1:237
- Statistical efficiency, 1:232
- Statistical graphics. *See* **Graphical statistical methods**
- Statistically significant result, 3:891
- Statistical matching, 3:828
- Statistical power
- hierarchical multiple regression, 3:998
 - inferential statistics, 2:459
 - likelihood ratio test, 2:534
 - null hypothesis significance testing, 2:697–698
 - one- versus two-tailed tests, 2:713
 - Page's L test, 2:722
 - sample size, 3:857–858
 - Shapiro-Wilk test for normality, 3:885
 - Type I error rate and, 2:646
 - Type II error, 3:1022
- Statistical significance**, 3:963–966
- binary decision task, 3:964
 - chance effect on empirical data, 3:963
 - complement of chance effects, 3:965
 - decision about, 3:966
 - implication of chance effects, 3:965
 - inferential statistics and, 3:964
 - research manipulation, 3:963–964
 - strength and meaningfulness versus, 3:1030
 - test statistic, 3:964
 - Type I and Type II errors, 3:966, 1019
 - See also* **Significance testing**
- Statistical Society of Canada, 1:27
- Statisticians' toolkit, 1:413–414
- Statistics
- definition, 2:457
 - measurement versus, 2:582, 583
- Statistics in Chemistry Award, 1:27
- Stecker, P., 1:208
- Steiger, Jim, 1:214
- Steinberg, Marlene, 3:977
- Stekloff, Vladimir A., 2:567
- Stem-and-leaf display**, 1:380, 413, 414, 3:966–967
- Stem cell research, 1:93
- Stemler, S. E., 2:484
- Stephan, Frederick W., 2:500
- Stephenson, William, 3:799
- Stephenson, W. Robert, 2:502
- Step-wise backward elimination, 2:558
- Sternberg, R. J., 1:411
- Sternberg, Robert, 1:404
- Stevens, James P., 2:670
- Stevens, S. S., 2:582
- Stevens, Stanley Smithy, 3:863
- Stochastic processes
- Markov chain and, 2:568
 - simulation experiments, 3:901
- Stochastic vector, 1:281
- Stock market, moving averages and, 2:637–638
- Stolzenberg, R. M., 1:59–60
- Stone-Romero, E. F., 3:996–998
- Story, M., 1:244
- Stout, William, 2:734
- Strange attractors, 1:373
- Strata, 3:812
- Stratified random sampling**, 3:967–971
- adaptive designs, 1:12
 - advantages, 3:970
 - cluster sampling versus, 1:151
 - defining strata, 3:970
 - definition, 3:967–968
 - example of, 3:968–969
 - quota sampling versus, 3:812

- simple random sampling versus, 3:968
- survey effort allocation, 3:969–970
- Strauss, Anselm, 1:418–420
- Strauss, H. J., 1:242–245
- Strength of relationship, 3:1030
- Strong, Edward K., Jr., 3:971
- Strong Interest Inventory (SII), 3:971–972**
- Strong Law of Large Numbers, 2:528
- Stroop, J. Ridley, 3:972
- Stroop Color and Word Test, 3:972**
- Structural equation modeling (SEM), 3:973–976**
 - attrition bias, 1:59
 - causal analysis, 1:124–125
 - data collection, 3:974
 - model estimation, 3:974–975
 - model evaluation, 3:975
 - model modification, 3:975–976
 - model specification, 3:973–974
 - overview of, 3:973
 - path analysis versus, 2:747
 - quasi-experiments, 3:807–808
 - sample size, 3:974
 - similar models, 3:973
 - software packages, 3:974, 975
 - tests of mediating effects, 3:996
- Structured Clinical Interview for DSM-IV Dissociative Disorders (SCID-D) Revised, 3:977
- Structured Clinical Interview for DSM-IV (SCID), 3:976–977**
- Structured interviews, 1:251
- Structure of Intellect model, 2:477
- Stuart, Alan, 1:108, 110
- Studentized range statistic, 3:1016
- Student's *t*. *See t test for two population means*
- Student Testing Program (STP), 1:45, 46
- Study, 2:658
- Stuetzle, W., 2:785
- Subject-specific models, 1:399–400
- Substantive population, 3:963
- Sufficient dimension reduction (SDR), 1:265–267
- Sugiura, Nariaki, 1:16
- Sullivan method, 1:11
- Summative achievement tests, 1:9
- Sum scores, 1:347
- Sunflower plot, 3:977–978**
- Supervised learning, 1:47, 267
- Support vector machines (SVM), 3:978–980**
 - data mining, 1:234
 - kernels, 3:979
 - principal component analysis, 2:786
 - soft margins, 3:979
 - software packages, 3:979
 - uses, 3:978–979
- Suppressor variable, 3:980–981**
- Survey of Consumer Finances, 2:663
- Survey of Income and Program Participation, 3:873
- Survey sampling, goal of, 2:689
- Survey weights, 2:690, 3:813, 981**
- Survival analysis, 3:982–987**
 - censoring and truncation, 3:982–983
 - Kaplan-Meier estimation, 3:983–985
 - log-rank tests, 3:985–986
 - semiparametric regression analysis, 3:986–987
- Survival function, 3:982
- Swain, George F., 2:500
- Swayne, D., 1:415
- Swineford, F., 2:546
- Symbolic interactionism, 1:418
- Symmetric difference distance, 1:283, 2:508
- Symmetry
 - Bowker procedure, 1:107–111
 - box and whisker plots, 1:111–112
- Syrus, 2:525
- SYSTAT, 1:358
- Systematic measurement error, 1:141, 2:585, 730, 3:944, 992–993
- System of Multicultural Pluralistic Assessment (SOMPA), 3:988**
- Systems Development Corporation, 2:608
- Szent-Györgyi, Albert, 2:425
- Tandem criteria of rotation, 1:174
- Tangible data, 1:220
- Tannery, Jules, 2:500
- Tanur, Judith M., 2:501
- Target, 1:13
- Tate, M. W., 1:152
- Tate, Robert, 1:95, 96, 97
- Tavana, M., 1:244
- Taxicab distance, 1:283
- Taxonomies, 1:196–197
- Taylor series expansions, 1:246–247
- Teachers
 - alternate assessment, 1:21–22
 - Incubation Model of Teaching, 3:1010
 - portfolio assessment, 2:776–777
- Teaching to the test, 2:438
- Technology
 - ability tests, 1:5
 - ecological momentary assessment, 1:297

- Technometrics*, 1:27, 3:1051
 Telephone interviews, 3:809
 Teller, A., 3:896
 Teller, E., 3:896
 Teller, Edward, 2:677
 Tellis, G. J., 1:385
 Temporal stability, 3:834
 Terciles, 2:756
 Terman, Lewis M., 1:1, 39, 203, 2:441, 476, 3:950–951
 TerraNova, 1:10
 Test, definition of, 1:8, 3:948
 Test alignment research, 1:182
 Test bias. *See* **Item and test bias**
 Test construction
 administration of test, 1:198
 content validity and, 1:182
 criterion-referenced tests, 1:197–199
 interpretation instructions, 1:198
 item response theory, 2:495
 items, 1:198–199
 Jackson and, 2:497
 naming the test, 1:198
 objectives represented by test, 1:198
 question construction, 3:809–811
 scoring instructions, 1:198
 self-report measures, 2:767–768
 true/false, 3:1013
 Test information, 2:495
 Test information function, 2:495
 Testing
 business of, 1:115
 ethical issues in, 1:318–321
 evaluation of, 1:115
 revisions and, 1:319
 Test length, 2:519, 3:935
 Test of English as a Foreign Language (TOEFL), 1:298
 Test of Memory and Learning, 3:846
Test-retest reliability, 3:992–995
 coefficient of, 3:993
 conditions for, 3:993–994
 determination of, 1:155
 error captured by, 3:994
 internal consistency and, 1:157
 reliability theory, 3:834
 standard error of measurement, 3:944
 Test Reviews Online, 1:115
 Test security, 1:173, 320–321
Tests in Print, 1:10, 115
Tests of mediating effects, 3:995–998
 Baron and Kenny hierarchical multiple regression procedure, 3:997–998
 causal model testing, 3:998
 model specification, 3:996–997
 research design, 3:995–996
 statistical approaches, 3:996
 Test standardization, 2:474
 Test statistic, 3:964
*Test Statistic*_{calculated}, 3:890
 Test variable, 3:990
 Testwise alpha, 1:104
Text analysis, 3:999
 Text mining, 1:235
Thematic Apperception Test (TAT), 2:769, 790–792, 3:1000
 Theology, Ockham's razor and, 2:707–708
 Theoretical sampling, 1:419
 Theory
 appraisal criteria, 1:420
 classical test, 1:140–143, 3:1014–1016
 decision, 1:236–240
 generalizability, 1:142, 204, 3:1016
 Gf-Gc intelligence, 1:403–406
 grounded, 1:418–421, 418–421
 item response, 2:493–496
 probability, 1:129, 2:457
 reliability, 3:834–835
 research and, 1:72–73
 signal detection, 3:886–889
 validity, 3:1032–1035
Theory of Games and Economic Behavior (Von Neumann & Morgenstern), 1:237
Theory of Mental Tests (Gulliksen), 1:142
 Theory of true and error scores. *See* **Classical test theory**
 Thermal cooling, simulation of, 3:896–900
 Thijs, H., 1:263
 Third variable problem, 3:937
 Thomason, N., 1:26
 Thompson, B., 2:549
 Thomson Corporation, 3:872
 Thorndike, Edward L., 2:441, 477, 3:1033
 Threats to validity, 3:805–806, 1035
Three-card method, 3:1000–1001
 Three-parameter logistic model, 2:493
 Three-stage least squares estimation, 2:471
 Three-way ANOVA, 1:32–33
 Thurstone, Louis L., 1:2, 53, 174, 335, 341, 2:477, 536–537, 3:862, 864, 1002
Thurstone scales, 3:1002–1005

- advantages/disadvantages, 3:1004–1005
- attitude tests, 1:53
- equal-appearing intervals, 3:1002–1003
- paired comparisons, 3:1003–1004
- semantic differential method, 3:880
- successive intervals, 3:1004
- Tibbetts, S. G., 3:1043
- Time
 - causality and, 1:124
 - data collection and, 1:221
 - Delphi technique and, 1:243
 - dynamic models, 3:901
 - human behavior and use of, 1:114–115
 - longitudinal/repeated measures data, 2:558–561
 - parameter invariance, 2:735
 - personal projects analysis, 2:763
 - single-subject designs, 3:904
- Time plot, 3:1005
- Time series analysis, 3:1005–1009**
 - ARIMA models, 3:1007–1009
 - autoregressive, integrated, moving averages (ARIMA)
 - modeling, 1: 64
 - chance, 1:130–131
 - components, 3:1005–1007
 - definition, 3:1005
 - diagrams, 1:412
 - errors, 1: 62
 - example of, 3:1006–1007
 - quasi-experiments, 3:806
 - software packages, 3:1009
 - time-series regression modeling, 1: 64
- Time-to-event data, 3:982–987
- Timm, Neil H., 2:501, 608
- Timothy W. v. Rochester, N.H., School Dist.*, 2:456
- Tindal, G., 1:21
- Title I, Elementary and Secondary Education Act, 1:19, 20
- Title IX, Educational Amendments (1972), 3:876
- Tobin, James, 3:831
- Tobit regression, 3:831
- Topinard, Paul, 1:36
- Torrance, E. Paul, 3:1009–1010**, 1011, 1012
- Torrance Center for Creative Studies, 3:1010
- Torrance Tests of Creative Thinking (TTCT)**, 3:1010–1011, 1012
- Torrance Thinking Creatively in Action and Movement (TCAM)**, 3:1010, 1011–1012
- Total scores, 1:347
- Trace, of matrix, 1:307
- Trace lines, 2:491
- Trade tests, 1:40
- Training data
 - classification and, 1:267
 - discriminant analysis and, 1:267, 268
- Training phase, 1:47
- Trait theory, 2:769
- Transparency, 1:232
- Traveling salesman problem, 3:899
- Treatment validity, 3:841
- Treatment variable. *See* **Independent variable**
- Tree diagram, 3:1012–1013**
- Trends, 1:328, 3:1005
- Trenkler, D., 2:515
- Trgovac, M., 2:430
- Triangulation, 3:870
- Tri-archic theory of intelligence, 1:404
- Trimean, 2:587, 591
- Trimmed mean, 1: 65, 2:587, 591
- True/false items, 3:1013–1014**
- True negatives, 1:255
- True positives, 1:255
- True random number generators, 3:815
- True score, 3:1014–1016**
 - assumptions underlying, 3:1015
 - definition, 1:141, 3:933
 - error and, 3:1014–1016
 - fallible versus, 1:130
 - reliability and, 2:730, 3:1015
 - theory of, 3:1014–1016
- True variance, 3:993
- Truncated data, 3:982, 983
- Trust, in research, 1:324
- Tsai, Chih-Ling, 1:16
- T scores, 3:945–946**, 954, **989**
- t* statistic, 3:964
- t test for two population means, 2:513, 3:990–992**
 - assumptions of, 3:990
 - computations, 3:991
 - curve from, 2:693–694
 - example of, 3:990–991
 - one-way analysis of variance, 2:713
 - paired samples, 2:723–726
 - partial regression coefficient, 2:738
 - research hypothesis, 3:990–991
 - sample size, 3:857–858
 - SPSS calculation of, 3:991–992
 - Wilcoxon Rank-Sum test versus, 2:565
 - Wilcoxon Signed Ranks Test versus, 3:1052
- Tucker, L. R., 3:849, 850
- Tufte, Edward R., 1:414

- Tukey, John W., 1:32, 111, 155, 328–329, 359–360, 413, 2:580, 583, 647, 681, 722, 778, 3:966, 1016
Tukey-Kramer procedure, 2:647, 3:1016–1018
Tunney, John, 1:114
Turkel, W. J., 1:166
Turnbull estimator, 3:984–985
Tuskegee Syphilis Study, 1:322, 2:465–466, 481
2 x 2 factorial design, 1:348–351
2 x 2 table, 1:254, 256–257
Two-factor theory, Spearman's, 2:477
Two-level hierarchical linear model, 2:433–434
Two one-sided tests (TOST) procedure, 1:314–315
Two-stage cluster sample, 1:150
Two-stage least squares (TSLS) estimation, 2:470–471
Two-step efficient GMM estimator, 1:391
Two-tailed tests. *See* **One- and two-tailed tests**
Two-way ANOVA, 1:32, 151, 380
Tyler, Ralph, 1:203
Type K format, 2:643
Type I Error, 3:1019
 complete independence hypothesis, 1:166–167
 definition, 2:458, 3:1019
 multiple comparisons and, 2:644–646
 Newman-Keuls test, 2:681
 number of tests and, 1:103
 Peritz procedure, 2:760–762
 statistical significance and, 3:966
 t test and, 2:513
Type II Error, 3:1019–1023
 common misconceptions, 3:1023
 complete independence hypothesis, 1:167
 conditions for, 3:1020–1021
 definition, 2:459, 3:1019
 failure to reject null hypothesis and, 1:314
 power, 3:1022
 probability and, 3:1021–1022
 research design, 3:1022–1023
 statistical significance and, 3:966
 synonyms for, 3:1021
Ulam, Stanislaw Marcin, 2:627
Ultimate sampling units, 1:150
Unadjusted correlation, 3:849
Unbiased estimator, 3:1025–1026
Unconditional model, 2:434
Underachievement, 1:42–43
Underlying factors, 1:332
Unidimensional, dichotomous IRT models, 2:494
Unified concept of validity, 3:1034–1035
Uniform differential item functioning, 2:490
Uniformly minimum variance unbiased estimators (UMVUE), 1:239
Union bound, 1:294
Unique factors, 1:332, 341
United States Bureau of Labor, Current Population Survey, 1:150
United States Bureau of the Census, 1:248, 2:527, 3:827–828, 873
United States Congress, 2:678
United States Department of Agriculture, National Resources Inventory, 1:150
United States Department of Commerce, 3:873
United States Department of Education, 1:19–20, 206, 2:502, 518, 3:840, 873
United States Department of Health and Human Services, 1:322, 2:427, 429, 466, 481
United States Department of Labor, 1:4, 3:873
United States Employment Service, 1:2
United States Office of Special Education, 1:417
United States Public Health Service, 2:481
United States Social Services Administration, 3:873
Unit information priors, 1:77
Universal Declaration of Human Rights (1948), 2:466
Universal Nonverbal Intelligence Test (UNIT), 3:1026–1027
Universals, theory of, 2:708
University Microfilms International (UMI), 1:24
University of Chicago tradition of social research, 1:418
University of Minnesota Press, 2:609
Unobserved variables. *See* Latent variables
Unrestricted factor analysis, 1:342
Unsupervised dimension reduction, 1:265
Unsupervised learning, 1:235
Unsystematic measurement error, 1:141, 3:944, 992–993
Unwin, A. R., 1:415
Users
 face validity and, 1:337–340
 Information System Interaction Readiness Scales, 2:464–465
 modeling information for, 1:228–231
Uspenskii, Ya. V., 2:567
Utilitarian tradition, ethics in, 1:323
Valentin, D., 1:284
Validation
 construct, 3:1035
 definition, 1:9
 standardized achievement tests, 1:10
 See also Validity

Validity

- achievement tests, **1:9**
- alternate assessment, **1:21–22**
- attrition bias, **1:58**
- autocorrelation, **1: 62**
- categorical, **1:254**
- concurrent, **1:200–201, 3:1029, 1033**
- consequences-of-testing, **3:949, 1034**
- content, **1:181–183, 3:949, 1034**
- convergent, **2:666–668**
- criterion, **1:200–202, 2:769, 3:949, 1033**
- definition of, **1:181, 3:1032**
- diagnostic, **1:254–258**
- discriminant, **2:666**
- ethics of testing selection and, **1:319**
- external, **3:806**
- face, **1:337–340, 2:759, 3:1034**
- internal, **3:805–806**
- performance-based assessment, **2:759**
- predictive, **1:200–201, 2:781–782, 3:1029, 1033**
- projective testing, **2:769**
- quasi-experiments, **3:805–806**
- reliability and, **1:178, 3:834, 1031**
- response process, **3:949**
- significance of, **3:1032**
- sources of, **3:949**
- stakeholders in, **3:1033**
- Standards for Educational and Psychological Testing* on, **3:949**
- statistical conclusion, **3:806**
- threats to, **3:805–806, 1035**
- unified concept of, **3:1034–1035**
- See also **Construct validity**; Validation;

Validity theory**Validity coefficient, 3:1029–1032**

- computation, **3:1030**
- concurrent validity and, **1:200**
- concurrent versus predictive evidence, **3:1029**
- criterion validity and, **3:1029, 1031, 1033**
- evaluating, **3:1031–1032**
- interpretation, **3:1030–1031**

Validity generalization, 3:1035**Validity theory, 3:1032–1035**

- constructivism and, **3:1035**
- future issues, **3:1035**
- history of, **3:1033–1034**
- philosophical challenges, **3:1035**
- validity in social context, **3:1032–1033**

See also **Diagnostic validity**

Values inventory, 2:532–533

Van Ness, John, **1:373**

Van Soest, J., **2:540**

Van Waveren Hudde, Johann, **1:86**

Vapnik, Vladimir, **3:978**

Vapnik Chervonenkis theory, **3:978**

Variable, 3:1035–1037

- categorical, **1:122**
- censored, **3:831**
- continuous, **1:183–185**
- criterion, **1:200, 249, 2:545**
- curse of dimensionality and, **1:209–211**
- definition of, **3:1035–1036**
- dependent, **1:249–250, 2:624, 3:1036**
- discrete, **1:184–185**
- endogenous, **2:468–469, 746, 3:974**
- estimator, **1:200, 202**
- exogenous, **2:468–469, 746, 3:974**
- group, **3:990**
- grouping, **3:990**
- independent, **2:453–454, 624, 3:1036**
- instrumental, **2:468–472**
- latent, **1:79, 332, 341, 345, 3:973**
- mediator, **2:624, 3:974, 995–998**
- moderator, **2:624–626**
- nominal, **1:122**
- ordinal, **1:122**
- outcome, **2:442**
- predicted, **2:624**
- predictor, **1:200, 202, 2:442, 545, 612, 624, 3:980**
- suppressor, **3:980–981**
- test, **3:990**
- types, **3:1036**

Variable deletion, 3:1037–1041

Variable-width box plot, **1:112**

Varian, H. R., **1:237–238**

Variance, 3:1042–1043

- heteroscedasticity and homoscedasticity, **2:431–432**
- homogeneity of, **2:442–443**

Variance inflation factors, **2:639–640**

Variogram, **3:925**

Vector ARMA models, **3:1009**

Vector correlation coefficients, **3:849**

Veltkamp, B. C., **1:166**

Venn, John, **1:129**

Verbal IQ, 3:1043–1044

Verbeke, G., **1:263**

Video compression algorithms, **1:230–231**

Vigotsky Test, **1:314**

Vineland Adaptive Behavior Scales II (VABS-II),
3:1044, 1045

Vineland Adaptive Behavior Scales (VABS),
3:1044, 1045

Vineland Social Maturity Scale, 3:1045

Violin plot, **1:112**

Visual processing, **1:3, 405**

Visual system, data compression for, **1:229–230**

Vital Statistics on American Politics, 3:873

Vocational instruments

 Armed Services Vocational Aptitude

 Battery, **1:3, 45**

 Career Assessment Inventory, **1:118**

 Career Development Inventory, **1:119**

 Career Maturity Inventory, **1:120**

 Differential Aptitude Test, **1:4, 260**

 Edwards Personal Preference Schedule, **1:299**

 Iowa Tests of Educational Development, **2: 488**

 Jackson Personality Inventory–Revised, **2:498**

 Jackson Vocational Interest Survey, **2:499**

 Kuder Occupational Interest Survey, **2:521**

 multiaptitude tests, **1:41**

 Occupational Information Network (O*NET)

 Ability Profiler, **1:4**

 Strong Interest Inventory, **3:971–972**

See also Job performance selection instruments

Volkobyskii, L. I., **2:567**

Von Braun, Werner, **1:217**

Von Mises, Richard, **2:457**

Von Neumann, John, **1:237, 2:627**

Voronoy, Georgy F., **2:568**

Wagner, T. H., **1:65**

Wald, A., **1:237**

Waldeck, Minna, **1:386**

Walker, Francis A., **2:500**

Walker-McConnell Scale of Social Competence and
 School Adjustment, **3:921**

Wallenstein, S., **1:154**

Walters, G. D., **2:430**

Warlick, Ken, **1:19**

Watson, James D., **1:90**

Wavelet analysis, **1:374–375**

Weak Law of Large Numbers, **2:528**

Weakliem, D. L., **1:17**

Wechsler, David, **2:475, 477, 757, 3:1043**

Wechsler Adult Intelligence Scale-III (WAIS-III),
 2:512, 3:1047, 1048

Wechsler Adult Intelligence Scale Revised (WAIS-R),
 2:640

Wechsler Adult Intelligence Scale (WAIS), 2:608,
 640, 757, 3:1047–1048

Wechsler-Bellevue Intelligence Scale, **2:477, 640**

Wechsler Individual Achievement Test-II (WIAT-II),
 3:1048

Wechsler Individual Achievement Test (WIAT),
 3:1048

Wechsler Intelligence Scale for Children-IV
 (WISC-IV), **3:1048**

Wechsler Intelligence Scale for Children-Revised
 (WISC-R), **3:988, 1047**

Wechsler Intelligence Scales, **3:1043**

Wechsler Memory Scale-III, **2:512**

Wechsler Preschool and Primary Scale of
 Intelligence (WPPSI-III), 3:1047, 1048, 1049

Wechsler Scales, **1:405, 2:478**

Wedderburn, R. W. M., **1:398**

Wegman, E., **2:727**

Weighted deviations model, **1:172**

Weighted GMM estimator, **1:390–391**

Weighted least squares (WLS) regression, **2:431–432,**
 530–532

Weighted mean, **2:587**

Weights

 assignment of, **2:518**

 internal consistency, **3:835**

 nonprobability sampling, **3:813**

 survey, **2:690, 3:813, 981**

Weisstein, Eric, **3:802**

Welch, Terry, **1:227**

Welch's t' statistic, **1:84–85**

Weldon, Raphael, **2:750**

Weldon, Walter Frank Raphael, **2:626–627**

Well-being, assessment of, **3:804–805**

Welsch, Roy, **2:681**

Weng, L.-J., **3:994**

Wesch, D., **3:819**

Western Educational Assessment, **1:120**

Western Psychological Services, **3:847**

West Haven-Yale Multidimensional Pain
 Inventory (WHYMPI), 3:1050–1051

Whitaker, D. J., **1:127**

Whitmore, Georgiana, **1:69**

Wilcox, R. R., **2:501, 593**

Wilcoxon, Frank, 3:1051

Wilcoxon Rank-Sum test, **2:565–567, 686, 3:1051**

Wilcoxon Signed Ranks Test, 3:1051, 1051–1053

Wilk, M. B., **3:884**

Wilkinson, Leland, **1:414**

Wilks's Lambda test, **2:671**

Willeke, M. J., **2:486**

William of Ockham, **2:707–709**

- Williams, W. M., 1:411
Willowbrook State School for the Retarded,
New York, 2:481
Wilson, David B., 2:596
Wimmer, R. D., 3:874
Window size, 3:865–866
Winsor, Charles P., 1:329
Winsteps, 3:822
Withdrawal designs, 3:905–906
Within-group interrater agreement, 1:66
Within-subjects ANOVA. *See* One-way
repeated measures ANOVA
Within-subjects designs. *See* **Longitudinal/repeated
measures data; Single-subject designs**
Witkin, H. A., 1:313
Wold, Herman, 2:740
Wold, Svante, 2:740, 744
Wolf, F. M., 1:301
Wolfe, Douglas, 2:686
Wood, J. M., 3:848
Woodcock, R. W., 2:475
Woodcock Johnson-III, 1:405, 2:478, 3:1053–1054
**Woodcock Johnson Psychoeducational Battery,
3:1053–1054**
Woodcock Language Proficiency Battery-Revised, 1:71
**Woodcock Reading Mastery Tests Revised
(WRMTR), 3:1054–1055**
Woodruff, David J., 1:157
World Book Company, 3:950
World Health Organization, 3:1044
World War I, 1:1
World War II, 1:2
Wright, B. D., 2:583
Wright, Sewall, 2:745
Wright, Steven, 2:503
Writing, grounded theory and, 1:419
w-scan, 3:866

 χ^2 distance, 1:281–282
Xerox University Microfilms, 1:25
X-ray cephalograms, 1:35

Yelton, Jack, 1:337
Yerkes, Robert, 2:476, 640
Yin, Z., 1:38
Ying, Z., 1:172
Y-intercept, 2:549
Yogo, J., 1:202
Yule, G. U., 3:894

Zaremba, Stanislaw, 2:568
Zeger, S. L., 1:400
Zeigler, L. H., 1:242–245
Zelinski, E. M., 1:187
Zero-reject approach, 2:456
Zimbardo Stanford Prison Study, 2:466
Zimmerman-Zumbo repeated measures ANOVA
on ranks, 1:381
Ziv, Jacob, 1:227
Zolotarev, Egor Ivanovich, 2:567
Zooming, 1:416
z scores, 3:945–946, 954, 989, 1057–1058
z statistical test, attrition bias and, 1:59
Zumbo, Bruno D., 2:501