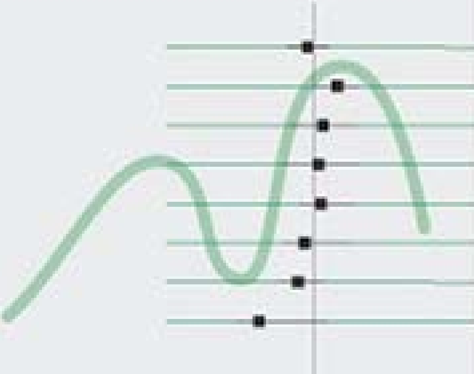
D.E. Matthews · V.T. Farewell

Using and Understanding Medical Statistics

4th, completely revised and enlarged edition

New Edition
2007



KARGER

Using and Understanding Medical Statistics

David E. Matthews · Vernon T. Farewell

Using and Understanding Medical Statistics

4th, completely revised and enlarged edition

42 figures and 113 tables, 2007

KARGER

Basel · Freiburg · Paris · London · New York ·
Bangalore · Bangkok · Singapore · Tokyo · Sydney

David Edward Matthews

BA, MA (Western Ontario), PhD (London);
Professor, Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, Ontario, Canada

Vernon Todd Farewell

B. Math, M. Math (Waterloo), PhD (London); Senior Scientist
MRC Biostatistics Unit, Cambridge, UK

Library of Congress Cataloging-in-Publication Data

Matthews, David E., 1948-
Using and understanding medical statistics / David E. Matthews Vernon T.
Farewell. – 4th, completely rev. and enl. ed.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978–3–8055–8189–9 (hard cover : alk. paper)
ISBN-10: 3–8055–8189–0 (hard cover : alk. paper)
1. Medical statistics. I. Farewell, Vernon T. II. Title.
[DNLM: 1. Biometry–methods. 2. Statistics. WA 950 M438u 2007]
RA409.M39 2007
610.1'5195–dc22 2006100514

Bibliographic Indices. This publication is listed in bibliograpahic services, including Current Contents® and Index Medicus.

Disclaimer. The statements, options and data contained in this publication are solely those of the individual authors and contributors and not of the publisher and the editor(s). The appearance of advertisements in the book is not a warranty, endorsement, or approval of the products or services advertised or of their effectiveness, quality or safety. The publisher and the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions or products referred to in the content or advertisements.

Drug Dosage. The authors and the publisher have exerted every effort to ensure that drug selection and dosage set forth in this text are in accord with current recommendations and practice at the time of publication. However, in view of ongoing research, changes in government regulations, and the constant flow of information relating to drug therapy and drug reactions, the reader is urged to check the package insert for each drug for any change in indications and dosage and for added warnings and precautions. This is particularly important when the recommended agent is a new and/or infrequently employed drug.

www.karger.com
Printed in Switzerland on acid-free paper by Reinhardt Druck, Basel
ISBN-13: 978–3–8055–8189–9 (hard cover : alk. paper), ISBN-10: 3–8055–8189–0 (hard cover : alk. paper)

To Nancy and Jane

Contents

Preface to the Fourth Edition . . . XI
Preface to the Third Edition . . . XV
Preface to the Second Edition . . . XVII
Preface to the First Edition . . . XIX

1 Basic Concepts . . . 1
1.1. Introduction . . . 1
1.2. Random Variables, Probability Distributions and Some Standard Notation . . 2
1.3. Characteristics of a Distribution: Mean, Median and Variance . . . 8

2 Tests of Significance . . . 11
2.1. Introduction . . . 11
2.2. An Example . . . 12
2.3. Common Features of Significance Tests . . . 17

3 Fisher's Test for 2 × 2 Contingency Tables . . . 19
3.1. Introduction . . . 19
3.2. Details of the Test . . . 19
3.3. Additional Examples of Fisher's Test . . . 25

4 Approximate Significance Tests for Contingency Tables . . . 28
4.1. Introduction . . . 28
4.2. The χ^2 Test for 2 × 2 Tables . . . 28
4.3. The χ^2 Test for Rectangular Contingency Tables . . . 35
4.4. Using Statistical Tables of the χ^2 Probability Distribution . . . 42

5 Some Warnings Concerning 2 × 2 Tables . . . 46
5.1. Introduction . . . 46
5.2. Combining 2 × 2 Tables . . . 46
5.3. Matched Pairs Binary Data . . . 51

6 Kaplan-Meier or 'Actuarial' Survival Curves 54
6.1. Introduction 54
6.2. General Features of the Kaplan-Meier Estimate 56
6.3. Computing the Kaplan-Meier Estimate 62
6.4. A Novel Use of the Kaplan-Meier Estimator 64

7 The Log-Rank or Mantel-Haenszel Test for the Comparison of Survival Curves 67
7.1. Introduction 67
7.2. Details of the Test 68
7.3. Evaluating the Log-Rank Test – A Simple Example 71
7.4. More Realistic Examples 74

8 An Introduction to the Normal Distribution 76
8.1. Introduction 76
8.2. Basic Features of the Normal Distribution 77
8.3. The Normal Distribution and Significance Testing 81
8.4. The Normal Distribution and Confidence Intervals 84
8.5. Using Normal Tables 87

9 Analyzing Normally Distributed Data 90
9.1. Introduction 90
9.2. Some Preliminary Considerations 91
9.3. Analyzing a Single Sample 96
9.4. Comparisons Based on the Normal Distribution 100
9.5. Testing the Equality of Variances 106

10 Linear Regression Models for Medical Data 111
10.1. Introduction 111
10.2. A Historical Note 112
10.3. Multiple Linear Regression 113
10.4. Correlation 118
10.5. The Analysis of Variance 122

11 Binary Logistic Regression 128
11.1. Introduction 128
11.2. Logistic Regression 129
11.3. Estimation in 2 × 2 Tables 131
11.4. Reanalysis of a Previous Example 135
11.5. The Analysis of Dose-Response Data 136

12 Regression Models for Count Data 141
12.1. Introduction 141
12.2. The Model for Poisson Regression 142
12.3. An Experimental Study of Cellular Differentiation 145

13 Proportional Hazards Regression 149
13.1. Introduction 149
13.2. A Statistical Model for the Death Rate 150
13.3. The Lymphoma Example 152
13.4. The Use of Time-Dependent Covariates 157

14 The Analysis of Longitudinal Data 160
14.1. Introduction 160
14.2. Liang-Zeger Regression Models 161
14.3. Random Effects Models 164
14.4. Multi-State Models 167

15 Analysis of Variance 174
15.1. Introduction 174
15.2. Representing Categorical Information in Regression Models 175
15.3. Understanding Two-Factor Interactions 181
15.4. Revisiting the INR Study 183

16 Data Analysis 191
16.1. Introduction 191
16.2. Quality Data 191
16.3. Initial or Exploratory Analysis 193
16.4. Primary Analysis 196
16.5. Secondary Analyses 200
16.6. A Formal Discussion of Multiple Comparisons 202
16.7. False Discovery Rates 206

17 The Question of Sample Size 207
17.1. Introduction 207
17.2. General Aspects of Sample Size Calculations 207
17.3. Two Examples of Sample Size Calculations 211
17.4. Some Hazards of Small Studies 214

18 The Design of Clinical Trials 217
18.1. Introduction 217
18.2. General Considerations 217
18.3. Trial Organization 220
18.4. Randomized versus Historical Controls 222
18.5. Intention to Treat 227
18.6. Factorial Designs 228
18.7. Repeated Significance Testing 228
18.8. Sequential Analysis 230

19 Further Comments Regarding Clinical Trials 234
19.1. Introduction 234
19.2. Surrogate Endpoints 234

19.3. Active Control or Equivalence Trials 236
19.4. Other Designs 237
19.5. Multiple Outcomes 239
19.6. Multiple Treatment Arms 242
19.7. Sequential Designs for Efficacy-Toxicity Trials 244
19.8. Stochastic Curtailment 250

20 Meta-Analysis 254
20.1. Introduction 254
20.2. Background 255
20.3. Study Heterogeneity 256
20.4. An Illustrative Example 257
20.5. Graphical Displays 260
20.6. Sensitivity 262

21 Epidemiological Applications 263
21.1. Introduction 263
21.2. Epidemiological Studies 263
21.3. Relative Risk Models 264
21.4. Odds Ratio Models 268
21.5. Confounding and Effect Modification 270
21.6. Mantel-Haenszel Methodology 273
21.7. Poisson Regression Models 275
21.8. Clinical Epidemiology 280

22 Diagnostic Tests 282
22.1. Introduction 282
22.2. Some General Considerations 282
22.3. Sensitivity, Specificity, and Post-Test Probabilities 286
22.4. Likelihood Ratios and Related Issues 292

23 Agreement and Reliability 298
23.1. Introduction 298
23.2. Intraclass Correlation Coefficient 300
23.3. Assessing Agreement 303
23.4. The Kappa Coefficient 304
23.5. Weighted Kappa 306
23.6. Measures of Agreement for Discrete Data 307
23.7. The Dependence of κ on Prevalence 308

References 310
Subject Index 314

Preface to the Fourth Edition

Twenty-five years have elapsed since we first began drafting our thoughts about using and understanding the statistical methods commonly employed in medical research. In the meantime, AIDS, SARS and avian flu, to name only three public health concerns, have burst onto the world stage. So, too, have fears of terrorism, the consequences of global warming and the next pandemic. Communication has been transformed, not least through the development of the World Wide Web. Twenty-five years ago, it would have required remarkable prescience to foretell, even in the vaguest of terms, how circumstances and events would unfold.

Even to predict the changes in attitudes that have occurred concerning smoking in public places, and the dangers of unprotected, excessive exposure to sunlight might have seemed highly speculative when *Using and Understanding Medical Statistics* first appeared in print.From the perspective of our goals in conceiving the first edition, the fairly widespread adoption of the phrase 'evidence-based medicine' is particularly noteworthy. It implicitly reflects the increased understanding, by physicians, that a basic grasp of statistical concepts and a passing appreciation for what statistical analysis can and cannot do is essential if one wants to be able to read and evaluate the medical literature.

When the first edition appeared, we hoped to make a small contribution to such effective use of the medical literature. Never in our wildest dreams did we imagine that four editions would be published in English, as well as foreign language editions in Italian, Spanish and in Japanese. And when we first began writing, we certainly did not anticipate the topics that would become commonplace in the medical research literature, and which we might therefore feel were appropriate to try and explain to our readers.

How does this edition differ from its predecessor? In our opinion, the fourth edition represents the most substantial revision of *Using and Understanding Medical Statistics* since the book was first published in 1984. As medical research has evolved, the statistical methods that are used to sift study data and adduce convincing evidence of improvement and innovation have become increasingly sophisticated. As a result, we have added entirely new chapters on Poisson regression, the analysis of variance, meta-analysis, diagnostic tests and the subject of measurement agreement and reliability. In addition, there are sections describing new topics in the chapters on longitudinal studies, data analysis, and clinical trials. Because statistical software is now widely available, we have removed the nine pages of statistical tables pertaining to Fisher's exact test; there are now many computational tools that will evaluate the exact significance level of this widely-used hypothesis test. Since there is now a chapter describing Poisson regression, we have also been able to add a new section to the chapter on epidemiological applications, one that describes the use of this tool to analyze the classic cohort study of smoking and coronary mortality reported by Doll and Hill. The changes in the public attitude towards smoking to which we previously referred are in large measure due to the pioneering efforts of Drs. Doll and Hill, and their work provides an outstanding example of fruitful collaboration between medical and statistical scientists. Finally, we must admit that the goal we identified in the first edition to have most chapters represent a single evening's reading has proved increasingly difficult to achieve.

First drafts of most of this new material were developed last year while DEM was an antipodean sabbaticant in the Centre for Clinical Epidemiology and Biostatistics at the University of Newcastle, and VTF made use of the excellent research facilities of his employer, the MRC Biostatistics Unit in Cambridge, England. We want to thank the Director and staff of the Centre – and particularly Professor Robert W. Gibberd – for generously providing DEM with a quiet office, as well as access to library and computing facilities. We also thank Professor Simon G. Thompson, the Director of the Biostatistics Unit, and the Unit staff for their support of VTF's efforts. Although the authors were half a world apart, the tangible encouragement of colleagues made writing and collaboration on a first draft of the fourth edition relatively easy.

Thanks are also due to our publisher, S. Karger AG in Basel, and especially to Rolf Steinebrunner, in production management, and Ms. Deborah Lautenschlager with whom we worked on this edition. We are particularly grateful to Rolf who has overseen the publication of all four editions of the book, and plans to retire at the end of 2006. It has been our privilege to enjoy such a long and fruitful relationship with Rolf, and everyone at Karger who has participated in publishing *Using and Understanding Medical Statistics*. We hope

Rolf's retirement years will be as fulfilling, and enjoyable, as his working career at Karger seems to us to have been.

For both of us, this year, 2006, marks significant anniversaries of our weddings. In recognition of those pivotal events in our personal lives, once again we dedicate whatever our efforts may have achieved to the two special individuals whose love and support have sustained and enriched our lives for more than half a century of unforgettable, shared experience.

D.E. Matthews
V.T. Farewell

Preface to the Third Edition

The world today is a very different place from what it was 12 years ago. Setting aside all the political changes that have occurred since *Using and Understanding Medical Statistics* was first published, there have been extraordinary developments in computing technology that have affected both the practice of medicine and the statistical methods that researchers use to evaluate medical progress. We will leave it to you, the readers of this book, to reflect on how medical practice has changed. From the statistical perspective, consider that when the first edition was published, access to statistical packages that would fit relative risk regression models to survival data was somewhat limited. More often than not, use of this software would have required the assistance of a statistician familiar with the input-output quirks and other vagaries of the routines. Now, the same models are a standard feature of many commercial software packages. Data input is often accomplished using spreadsheet editors, and most of the key aspects of model selection and fitting can be carried out by simply clicking the mouse on the correct item in a pull-down menu. Seconds later, the results of the analysis scroll across the screen. Perhaps the most astonishing aspect of this revolution in statistical analysis is the fact that one can carry out such analyses virtually anywhere – sitting in an airplane, riding on a train, or logged in from an office on one continent to a remote machine halfway round the globe.

The ease with which statistical analyses can now be carried out makes the focus of this book all the more important. Readers of the medical literature in former times may well have thought that a statistically significant result was, perhaps, the primary determinant in the editorial process of selection and publication. However, it has not been, and is not, sufficient simply to carry out some statistical analysis of medical data with this goal in mind.

Now that complex statistical analyses are easy to execute, it is particularly important that the focus for medical researchers shifts from computation to interpretation and understanding. Readers of the medical journals, now more than ever, need the ability to read and critically appraise the results of various studies that deal with issues in the practice of medicine such as new treatments for a specific disease, the natural history of conditions such as AIDS, or the public health benefit of a new screening program for breast or prostate cancer.

What is different about the third edition of *Using and Understanding Medical Statistics?* First, there are two new chapters. One of these provides readers with an introduction to the analysis of longitudinal data. We describe two quite different approaches to the analysis of such studies; both methods are beginning to find their way into the mainstream of medical literature. The second new chapter augments material concerning the design of clinical trials that appeared in the first and second editions. Readers are introduced to topics such as the use of surrogate markers, multiple outcomes, equivalence trials, and the design of efficacy-toxicity studies.

In addition to these new chapters, we have reorganized the last third of the book so that the actual order in which topics are introduced precedes their routine use. In this respect the second edition contained one glaring pedagogical error that we are pleased to have the opportunity to rectify. As well, we have taken great pains to carefully re-phrase sentences and paragraphs that did not stand up to scrutiny. We were greatly assisted in this exercise by a Danish medical statistician, Dr. Jørgen Hilden, who sent us several pages of constructive remarks that revealed deficiencies in the second edition. We, and our readers, owe Dr. Hilden a substantial debt of gratitude for this generous exercise of his experience, insight, and labour.

We conclude with one final observation concerning the impact of technological change on normal patterns of work. When the first edition was being prepared, Mrs. Joy Hoggarth typed the manuscript for us, and did a superb job. The second edition was typeset by Ms Lynda Clarke, whose productivity at the keyboard was astonishing. Now both of us write directly onto computers or X-terminals that occupy a corner of the desk, so we have no one to thank for capably assisting us in the preparation of the manuscript for the third edition. On the other hand, we are pleased to say that no technological change prevents us from dedicating this third edition to the same very special people to whom the first and second editions were also dedicated.

D.E. Matthews
V.T. Farewell

Preface to the Second Edition

Slightly less than four years have elapsed since the preface to the first edition was written. In the meantime, we have been surprised, and pleased, by the response to the first edition. The letters and comments which we have received from readers and reviewers on several continents have brought us much satisfaction. Suggestions and criticisms have helped us to understand specific topics where the first edition failed to meet the goals which we had established. Despite our best intentions, there were inevitable errors in the first edition which we were anxious to correct. Consequently, when the publisher inquired about the possibility of a revised edition, we realized that it would be an opportunity to rectify both kinds of flaws simultaneously.

How do the two editions differ? Apart from minor corrections to Table 3.4 and the elimination of errors which appear to be randomly distributed through the chapters, the principal differences may be found in the second half of the book. The example in chapter 10 has been changed to one which we believe suits better the purpose we intend to achieve. Sections have been added to chapters 11, 12 and 14 which treat topics that were previously omitted. In some ways, these additions reflect the changing face of medical statistics, and the clinical investigations in which statistical methods play an important role. However, the major difference between the editions is the addition of chapter 16, which concerns epidemiological studies. The topics treated in the final chapter illustrate just how much the use of sophisticated statistical analysis has permeated the recent practice of epidemiology. At the same time, this new chapter knits together the fabric of the book, drawing on methods which we have introduced in previous chapters to analyze data from various epidemiological studies. In that respect, chapter 16 does what no chapter in the first edition was able to do. We hope its inclusion in the second edition will help all readers, even those whose main interest is not directed towards epidemiology, to integrate their understanding and extend their appreciation for the use of statistical methods in medical research.

We are grateful to Ms Lynda Clarke in the Department of Statistics and Actuarial Science at the University of Waterloo. Her skill and cheerful co-operation made light work of all our changes in the process of preparing the revised manuscript.

D.E. Matthews
V.T. Farewell

Preface to the First Edition

The origins of this book can be traced to a short course which was offered in the autumn of 1980 to medical researchers at the Fred Hutchinson Cancer Research Center. The syllabus for that course was drawn up to meet the specific needs of those researchers. After re-examining the material we had presented, we felt that the content of the course and the approach we had adopted were different from their counterparts in familiar introductory books on statistics, even those which were written specifically for medical researchers. Unsolicited comments from course participants encouraged us to develop and expand our approach instead of filing the assorted tables and handouts. And so, through additions, deletions and numerous revisions, the final product haltingly took shape and assumed its present form.

Our aim now, as in 1980, is quite simple: to describe the statistical methodology which frequently is found in published medical research, particularly those papers concerned with chronic diseases. This presentation introduces, in some detail, fundamental statistical notions which are common to nearly every method of analyzing data – for example, significance tests. From these foundations we direct our attention to more advanced methods of analysis. In order to avoid excessive complexity in the initial chapters, we rely on the promise of bigger and better things to come to motivate the selected topics we choose to discuss. Nonetheless, there is sufficient introductory detail that we feel obliged to ask our readers to exercise more patience and endurance than most introductions to medical statistics require. We are convinced that solid beginnings are essential to any *useful* discussion of the important, more advanced methodology which frequently is used in modern medical research.

We have written for the motivated reader who is willing to invest a little time and effort in understanding statistical methods. On the other hand, our constant goal has been to write a book which could be read fairly easily in installments. We hope that most chapters represent a single evening's reading. Although one might decide to devote a little more time to some of the details, it should then be possible to tackle the next chapter. We shall be pleased if we have succeeded in achieving this goal; however, we do not wish to be regarded as competing with alternative evening reading which may be more interesting or exciting!

Except, perhaps, for an over-emphasis on chronic diseases, we believe that a medical student who understands the contents of this book will be well-informed regarding medical statistics. Whether medical students, who often regard statistics as an unnecessary evil, should and can be adequately motivated to master this material is an open question. We have not attempted to provide this motivation ourselves. In our view, the most persuasive arguments on behalf of the subject will always be those advanced by medical researchers who have themselves established a use for statistical analysis which does not depend on the editorial policy of medical journals.

The final preparation of this manuscript took place while one of us (V.F.) was visiting the Department of Biomathematics at the University of Oxford, and the other (D.M.) was visiting the Department of Medical Statistics and Epidemiology at the London School of Hygiene and Tropical Medicine. We want to thank Professors Peter Armitage and Michael Healy for making these visits possible. We are also greatly indebted to Mrs. Joy Hoggarth at the Fred Hutchinson Cancer Research Center for her superb preparation of the manuscript. She was greatly handicapped by being more than 5,000 miles from the authors.

An early version of this book was read by Dr. G.J. D'Angio of the Children's Hospital of Philadelphia; his helpful comments and criticisms had a significant influence on the final manuscript. We thank him for this and, in general, for his unwavering support of statistics in medical research. It is our hope that this book will help other investigators to develop a similar appreciation for the value of medical statistics.

D.E. Matthews
V.T. Farewell

1

Basic Concepts

1.1. Introduction

A brief glance through almost any recently published medical journal will show that statistical methods are playing an increasingly visible role in modern medical research. At the very least, most research papers quote (at least) one 'p-value' to underscore the 'significance' of the results which the authors wish to communicate. At the same time, a growing number of papers are now presenting the results of relatively sophisticated, 'multi-factor' statistical analyses of complex sets of medical data. This proliferation in the use of statistical methods has also been paralleled by the increased involvement of professionally trained statisticians in medical research as consultants to and collaborators with the medical researchers themselves.

The primary purpose of this book is to provide medical researchers with sufficient understanding to enable them to read, intelligently, statistical methods and discussion appearing in medical journals. At the same time, we have tried to provide the means for researchers to undertake the simpler analyses on their own, if this is their wish. And by presenting statistics from this perspective, we hope to extend and improve the common base of understanding which is necessary whenever medical researchers and statisticians interact.

It seems obvious to us that statisticians involved in medical research need to have some understanding of the related medical knowledge. We also believe that in order to benefit from statistical advice, medical researchers require some understanding of the subject of statistics. This first chapter provides a brief introduction to some of the terms and symbols which recur throughout the book. It also establishes what statisticians talk about (random variables, probability distributions) and how they talk about these concepts (standard notation). We are very aware that this material is difficult to motivate; it seems so distant from the core and purpose of medical statistics. Nevertheless, 'these dry bones' provide a skeleton which allows the rest of the book to be more precise about statistics and medical research than would otherwise be possible. Therefore, we urge the reader to forbear with these beginnings, and read beyond the end of chapter 1 to see whether we do not put flesh onto these dry bones.

1.2. Random Variables, Probability Distributions and Some Standard Notation

Most statistical work is based on the concept of a random variable. This is a quantity that, theoretically, may assume a wide variety of actual values, although in any particular realization we only observe a single value. Measurements are common examples of random variables; take the weights of individuals belonging to a well-defined group of patients, for example. Regardless of the characteristic that determines membership in the group, the actual weight of each individual patient is almost certain to differ from that of other group members. Thus, a statistician might refer to the random variable representing the weight of individual patients in the group, or population of interest. Another example of a random variable might be a person's systolic blood pressure; the variation in this measurement from individual to individual is frequently quite substantial.

To represent a particular random variable, statisticians generally use an upper case Roman letter, say X or Y. The particular value which this random variable represents in a specific case is often denoted by the corresponding lower case Roman letter, say x or y. The probability distribution (usually shortened to the distribution) of any random variable can be thought of as a specification of all possible numerical values of the random variable, together with an indication of the frequency with which each numerical value occurs in the population of interest.

It is common statistical shorthand to use subscripted letters – x_1, x_2, …, x_n, for example – to specify a set of observed values of the random variable X. The corresponding notation for the set of random variables is X_i, i = 1, 2, …, n, where X_i indicates that the random variable of interest is labelled X and the symbols i = 1, 2, …, n specify the possible values of the subscripts on X. Similarly, using n as the final subscript in the set simply indicates that the size of the set may vary from one instance to another, but in each particular instance it will be a fixed number.

Subscripted letters constitute extremely useful notation for the statistician, who must specify precise formulae which will subsequently be applied in particular situations which vary enormously. At this point it is also convenient to introduce the use of Σ, the upper case Greek letter sigma. In mathematics, Σ represents summation. To specify the sum $X_1 + X_2 + X_3$ we would simply write $\sum_{i=1}^{3} X_i$. This expression specifies that the subscript i should take the values 1, 2 and 3 in turn, and we should sum the resulting variables. For a fixed but unspecified number of variables, say n, the sum $X_1 + X_2 + \ldots + X_n$ would be represented by $\sum_{i=1}^{n} X_i$.

A set of values $x_1, x_2, \ldots, x_n$ is called a sample from the population of all possible occurrences of X. In general, statistical procedures which use such a sample assume that it is a random sample from the population. The random sample assumption is imposed to ensure that the characteristics of the sample reflect those of the entire population, of which the sample is often only a small part.

There are two types of random variables. If we ignore certain technicalities, a discrete random variable is commonly defined as one for which we can write down all its possible values and their corresponding frequencies of occurrence. In contrast, continuous random variables are measured on an interval scale, and the variable can assume any value on the scale. Of course, the instruments which we use to measure experimental quantities (e.g., blood pressure, acid concentration, weight, height, etc.) have a finite resolution, but it is convenient to suppose, in such situations, that this limitation does not prevent us from observing any plausible measurement. Furthermore, the notation which statisticians have adopted to represent all possible values belonging to a given interval is to enclose the end-points of the interval in parentheses. Thus, (a, b) specifies the set of all possible values between a and b, and the symbolic statement '$a < X < b$' means that the random variable X takes a value in the interval specified by (a, b).

The probability distribution of a random variable is often illustrated by means of a histogram or bar graph. This is a picture which indicates how frequently each value of the random variable occurs, either in a sample or in the corresponding population. If the random variable is discrete, the picture is generally a simple one to draw and to understand. Figure 1.1a shows a histogram for the random variable, S, which represents the sum of the showing faces of two fair dice. Notice that there are exactly 11 possible values for S. In contrast to this situation, the histogram for a continuous random variable, say systolic blood pressure, X, is somewhat more difficult to draw and to understand. One such histogram is presented in figure 1.1b. Since the picture is intended to show both the possible values of X and also the frequency with which they arise, each rectangular block in the graph has an area equal to the propor-

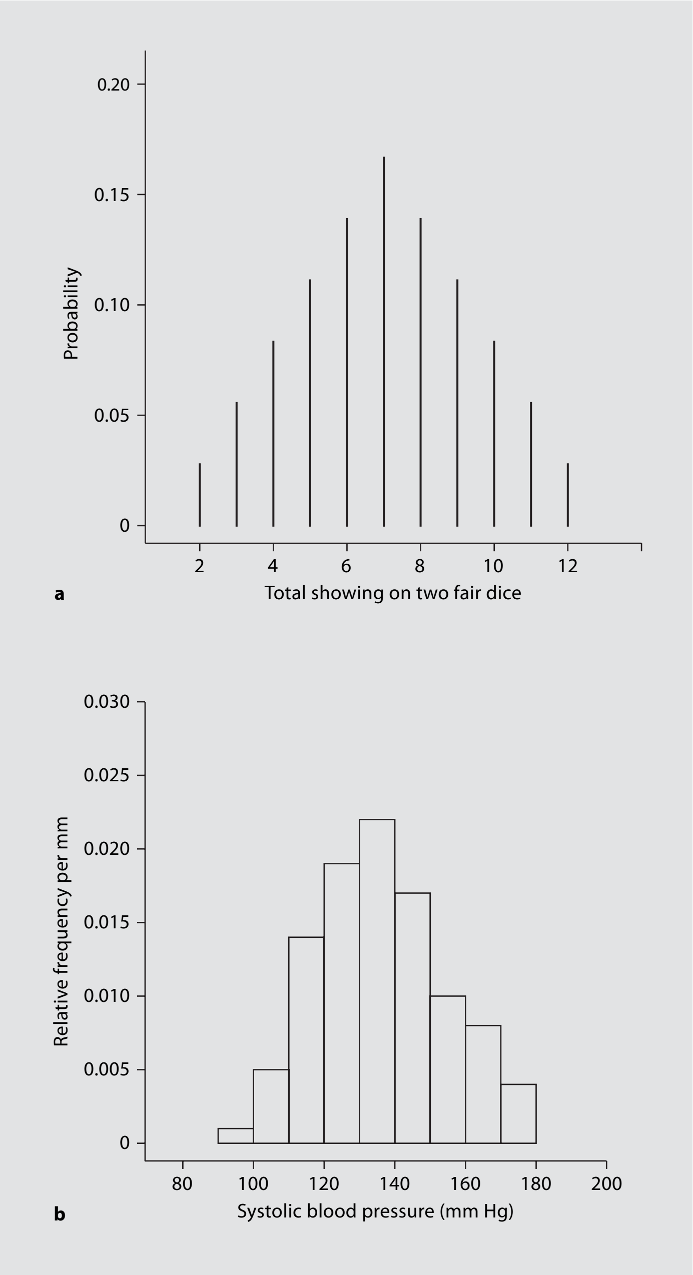


Fig. 1.1. Histograms of random variables. **a** The discrete random variable, S, representing the sum of the showing faces for two fair dice. **b** One hundred observations on the continuous random variable, X, representing systolic blood pressure.

tion of the sample represented by all outcomes belonging to the interval on the base of the block. This has the effect of equating frequency, or probability of occurrence, with area and is known as the 'area = probability' equation for continuous random variables.

To a statistician, histograms are simply an approximate picture of the mathematical way of describing the distribution of a continuous random variable. A more accurate representation of the distribution is obtained by using the equation of a curve which can best be thought of as a 'smooth histogram'; such a curve is called a probability density function. A more convenient term, and one which we intend to use, is probability curve.

Figure 1.2a shows the probability curve, or smooth histogram, for the continuous random variable, X, which we used above to represent systolic blood pressure. This curve is, in fact, the probability curve which has the characteristic shape and equation known as a 'normal distribution'. Random variables that have a normal distribution will recur in subsequent chapters, and we intend to explain their properties and uses in more detail at that time. For the present, however, we want to concentrate on the concept of the area = probability equation. Figure 1.2b shows two shaded areas. One is the area below the curve and above the interval (110, 130). Recall that the symbol (110, 130) represents all blood pressure measurements between 110 and 130 mm Hg. Because of the area = probability equation for the continuous random variable X, the shaded area above (110, 130) corresponds, pictorially, to the probability that systolic blood pressure in the population is between 110 and 130 mm Hg. This area can be calculated mathematically, and in this particular example the value is 0.323. To represent this calculation in a symbolic statement we would write $\Pr(110 < X < 130) = 0.323$; the equation states that the probability that X, a systolic blood pressure measurement in the population, is between 110 and 130 mm Hg is equal to 0.323.

The second shaded area in figure 1.2b is the area below the probability curve corresponding to values of X in the interval $(165, \infty)$, i.e., the probability that a systolic blood pressure measurement in the population exceeds 165 mm Hg. By means of certain calculations we can determine that, for this specific example, the probability that systolic blood pressure exceeds 165 mm Hg is 0.023; the concise mathematical description of this calculation is simply $\Pr(X > 165) = 0.023$.

Although the probability curve makes it easy to picture the equality of area and probability, it is of little direct use for actually calculating probabilities since areas cannot be read directly from a picture or sketch. Instead, we need a related function called the cumulative probability curve. Figure 1.3 presents the cumulative probability curve for the normal distribution shown in figure 1.2a. The horizontal axis represents the possible values of the random

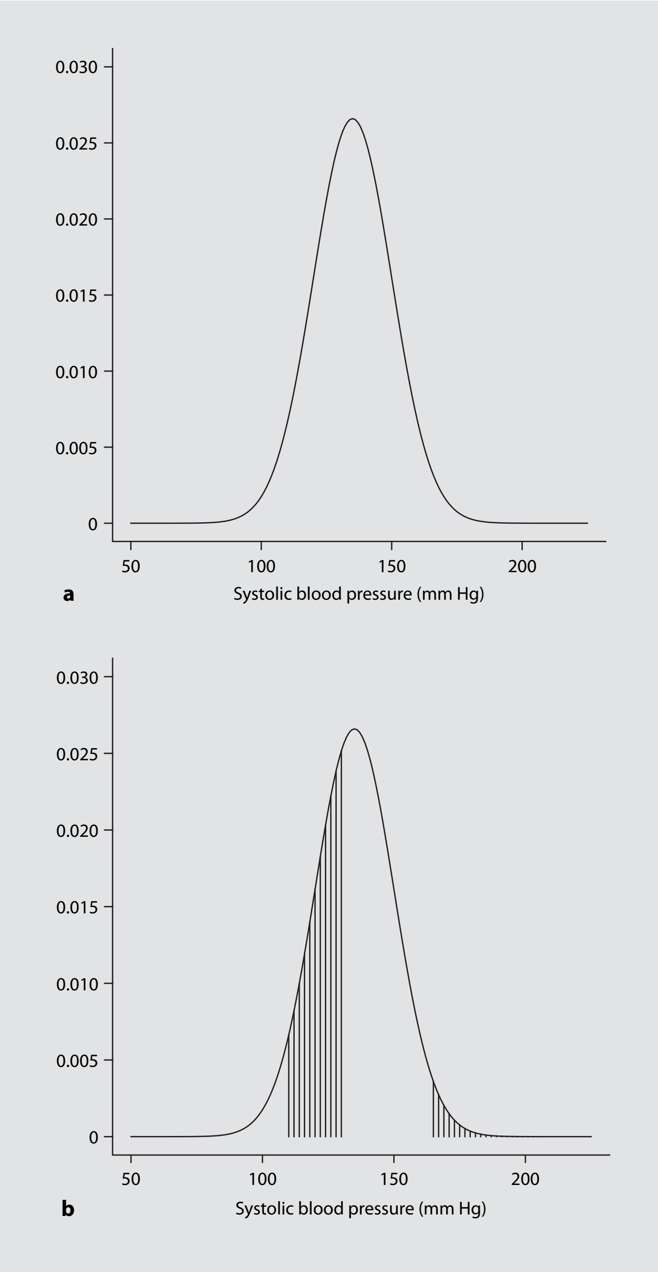


Fig. 1.2. A probability curve for the continuous random variable, X, representing systolic blood pressure. **a** As a smooth histogram. **b** With shaded areas corresponding to $\Pr(110 < X < 130)$ and $\Pr(X > 165)$.

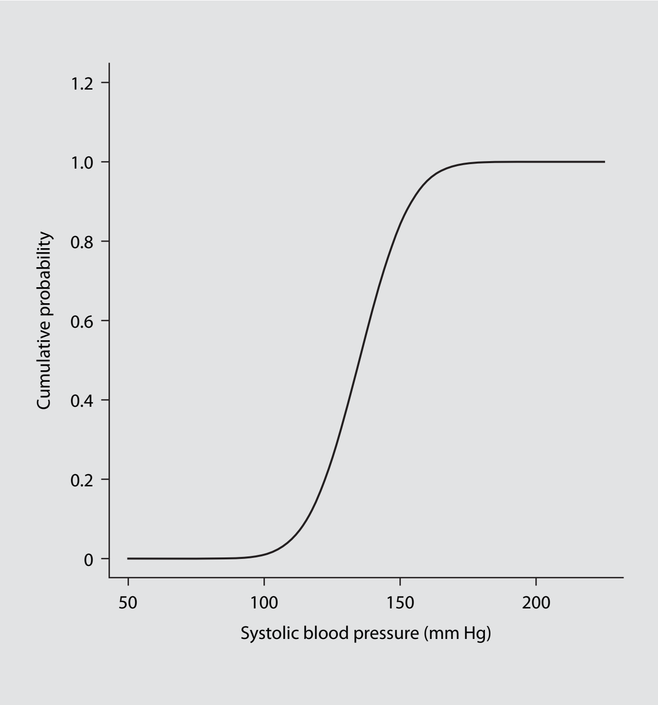


Fig. 1.3. The cumulative probability curve for the random variable, X, representing systolic blood pressure.

variable X; the vertical axis is a probability scale with values ranging from zero to one. The cumulative probability curve specifies, for each value a, say, on the horizontal axis, the probability that the random variable X takes a value which is at most a, i.e., $\Pr(X \leq a)$. This probability is precisely the area below the probability curve corresponding to values of X in the interval $(-\infty, a)$. In particular, if $a = \infty$, i.e., $\Pr(-\infty < X < \infty)$, the value of the cumulative probability curve is one, indicating that X is certain to assume a value in the interval $(-\infty, \infty)$. In fact, this result is a necessary property of all cumulative probability curves, and is equivalent to the statement that the area under any probability curve is always equal to one.

Clearly, a cumulative probability curve is more useful than the corresponding probability curve for actually calculating probabilities. For example, since the area under the probability curve always equals one, it follows that

$$\Pr(X > a) = 1 - \Pr(X \leq a).$$

Thus, the cumulative probability curve can be used to calculate the probability corresponding to the interval (a, ∞). And for the interval (a, b), where a and b are two specific values, it is fairly easy to show that

$$\Pr(a < X < b) = \Pr(X < b) - \Pr(X \leq a);$$

this is simply a difference of the cumulative probability curve evaluated at the two points a and b.

Comment:

Most readers have probably noticed that §1.2 did not contain any specific formulae for calculating probabilities, particularly in the case of continuous random variables. The reason for this is simple. More frequently than not, the calculations are sufficiently formidable, numerically, that statisticians have prepared standardized tables to make the evaluation of probabilities relatively simple. These tables, which often are tabulations of the cumulative probability curve, are generally called statistical tables; the values which we quoted for the normal distribution shown in figure 1.2b were obtained from a statistical table for the normal distribution. In subsequent chapters, as various common probability distributions arise in the exposition, we will discuss how to use the relevant statistical table, or the results of probability calculations provided by a statistical software package.

1.3. Characteristics of a Distribution: Mean, Median and Variance

If we wish to be absolutely precise about the distribution of a certain random variable, say X, then we must specify the equation of the probability curve if it is a continuous random variable, or the values of X and the probability with which they occur if X is a discrete random variable. However, if we only wish to indicate something about the distribution of X in general terms, it often suffices to specify the location and spread of the distribution of X. In common, everyday terms this is equivalent to the two usual answers to the question ‘Where is X’s office?’. If you wish to be precise you would answer ‘X’s office is Room 703 at 1024 Columbia Street, in the Cabrini Tower’. On the other hand, if you were only planning to offer a general answer to the question you would reply ‘X’s office is on Capitol Hill’.

1.3.1. Location

A measure of location for the distribution of a random variable, X, should tell us the value which is roughly central, in some sense, to the range of values where X is regularly observed to occur. The most common measure of location

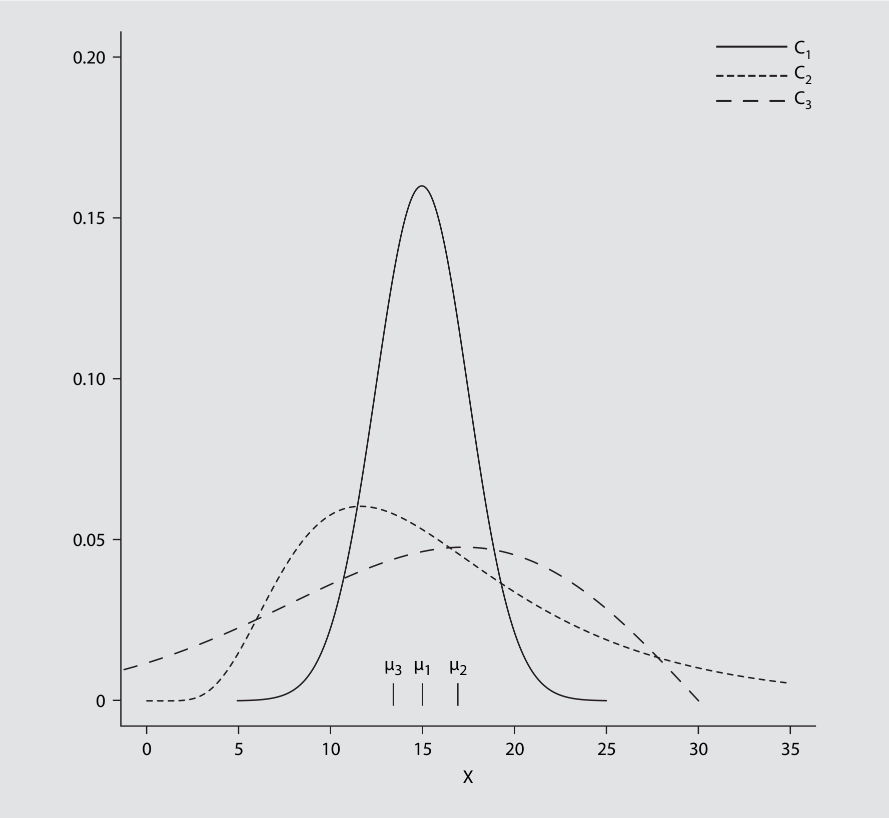


Fig. 1.4. Three probability curves, C_1, C_2 and C_3, having the same median (15) but different means μ_1, μ_2 and μ_3, respectively.

used by statisticians is the mean or expected value of X; this is denoted by the symbol E(X). If we think, for a moment, of a probability associated with the distribution of X as mass, then the mean value, E(X), is located at the center of mass. The Greek letter μ is often used to represent E(X).

A second measure of location is the median value of X. Most readers will recognize the median as the measure of location which is commonly used in the medical literature. The median of X is that particular value which equally divides the probability in the distribution of X, i.e., with probability 1/2 an observed value of X will exceed the median and with probability 1/2 an observed value will not exceed the median.

In general, the median of X and the mean, E(X), are different because the median is less influenced by extreme values of X which might occur than is

E(X). However, if the distribution of X below the median is a mirror image of the upper half of the distribution, then the distribution is said to be symmetric and the median and mean will coincide. To illustrate some of the differences between the mean and the median, figure 1.4 shows three probability curves which have the same median but different means.

1.3.2. Spread

The dispersion or spread of a distribution indicates how variable the quantity represented by X is expected to be from one observation to the next. The most common measure of spread for a random variable, X, is called the variance of X.

To define variance we begin with the constructed variable $(X - \mu)$, where $\mu = E(X)$; recall that E(X) is the mean of X and indicates, roughly, the center of the distribution. The constructed variable $(X - \mu)$ measures both the direction and the amount by which X deviates from its mean value μ. For the purposes of measuring spread in the distribution of X, the direction of this deviation is of less importance than the magnitude; however, large deviations in either direction influence the spread of the distribution more than small deviations. Therefore, we use $(X - \mu)^2$, the square of our constructed random variable, to indicate spread and call $E\{(X - \mu)^2\}$, the expected value of this random variable, the variance of the distribution of X. For convenience, the variance is frequently represented by the Greek symbol σ^2.

The square root of the variance, which is usually indicated by the symbol σ, is called the standard deviation of the distribution of X. By specifying the mean, μ, and standard deviation, σ, for the distribution of X, we are providing an approximate or general description of the entire distribution. But to evaluate probabilities for a distribution, or to perform other calculations, we need more specific information than just the mean and standard deviation. For accurate calculations we also require the equation of the cumulative probability curve or suitable tables. However, these are details which we intend to discuss in later chapters.

Too often the beginnings of any subject will kill the enthusiasm of the most eager reader. However, now that the groundwork has been done, we can proceed to discuss some of the basic procedures which statisticians have developed. The first of these which we will consider, and therefore the subject of chapter 2, is called a test of significance.

2

Tests of Significance

2.1. Introduction

Among the statistical methods employed in medical research, tests of significance enjoyed a pre-eminent position throughout most of the twentieth century. Of late, the focus has shifted towards estimation methods, a theme that we will first introduce in chapter 6, and subsequently encounter in chapters 8 and 9. Nonetheless, tests of significance are still widely used in the medical literature, and involve concepts related to estimation as well. Therefore, in this chapter, our goal is to identify the features which are common to all tests of significance to provide a basis for the topics we will consider in subsequent chapters.

A test of significance is a statistical procedure by which one determines the degree to which collected data are consistent with a specific hypothesis which is under investigation. There is a sense in which the process of arriving at a medical diagnosis may be compared to a test of significance. Consider the general practitioner treating a patient who has recently become ill. The patient's physical appearance and brief description of the symptoms prompt the physician to compile a list of perhaps three or four possible causes; each cause is a hypothesis of interest. Next, the physician uses knowledge and experience to examine the patient. The examination results in observed data concerning the patient's present condition. Finally, the physician evaluates the situation, determining the degree to which the observed data are consistent with each possible cause. The diagnosis identifies the single cause, or hypothesis, to which the weight of medical evidence points.

As we mentioned before, a significance test is quite similar to the diagnostic process described above. Of course, there are differences which are important. Nevertheless, the physician and the statistician are each interested in assessing the degree to which observed data are consistent with a specific hypothesis. In statistical parlance, the result of the evaluation process is called

Table 2.1. A 2 × 2 contingency table summarizing the data described in Thomas et al. [1]

	Field size too small	Field size adequate	Total
Operative site relapse	2	2	4
No operative site relapse	21	234	255
Total	23	236	259

the significance level ('p-value') of the data with respect to the hypothesis. Of course, there are assumptions and calculations which are an implicit part of the evaluation process. In fact, there are six features which are common to all significance tests, and §2.3 is chiefly a discussion of these features. But first, we propose to consider a specific example, and to explain the assumptions and arguments which give rise to the significance level of a test and the resulting conclusion.

2.2. An Example

Thomas et al. [1] discuss relapse patterns in irradiated Wilms' tumor patients. In particular, they describe a review of radiotherapy records which indicated that 23 out of 259 patients were judged to have been treated with a radiation field of inadequate size. An obvious hypothesis of interest is the suggestion that inadequate radiotherapy is associated with a higher risk of relapse in the operative site. Subsequent investigation of these 259 patients revealed that there were four relapses in the operative site, two of which occurred in the 23 patients who received inadequate radiotherapy. The data may be concisely presented in a summary table containing two rows and two columns (see table 2.1); tables of this form are usually called 2 × 2 contingency tables.

A major purpose of this study of 259 patients was to determine whether inadequate radiotherapy is associated with a higher risk of relapse in the operative site. If we examine the data, we see that the relapse rate among patients whose field size was adequate was 2/236 (0.9%), whereas the corresponding figure for patients who received inadequate radiotherapy was 2/23 (8.7%). If Thomas et al. were only interested in the hypothesis concerning relapse in the operative site for the 259 patients studied, then the data indicate that inadequate field size is associated with an elevated relapse rate in those 23 patients. However, this is not the question which the researchers want to answer. They are interested in the population of patients who are treated for Wilms' tumor.

Table 0

	Field size too small	Field size adequate	Total
Operative site relapse	0	4	4
No operative site relapse	23	232	255
Total	23	236	259

Table 1

1	3	4
22	233	255
23	236	259

Table 2

2	2	4
21	234	255
23	236	259

Table 3

3	1	4
20	235	255
23	236	259

Table 4

4	0	4
19	236	255
23	236	259

Fig. 2.1. The five possible 2 × 2 tables with fixed marginal totals 4, 255, 23 and 236.

The 259 patients in the study constitute a sample from that population. On the basis of the data collected from the sample, Thomas et al. wish to infer whether the relapse rate in the operative site is higher for all patients in the population who receive inadequate radiotherapy.

In order to arrive at any conclusions regarding the hypothesis of an elevated risk of relapse based on the sample data, we must assume that the sample is truly representative of the population in all pertinent respects. However, even if we assume that this is true, how should the results of the study be interpreted? Should we conclude that the sample relapse rates (0.009 and 0.087) are the same as the population relapse rates, and therefore inadequate field size is associated with a ten-fold increase in the risk of relapse? Surely not! After all, the sample size is fixed at 259 patients, 23 of whom received inadequate radiotherapy, and only four relapses in the operative site were observed. If we adjust the numbers in the four cells of the 2 × 2 table, always taking care to maintain the same row and column totals, then we see that only five different 2 × 2 tables

Table 2.2. Sample relapse rates in tables 0 through 4 of figure 2.1

	Table #				
	0	1	2	3	4
Field size too small	0.0	0.044	0.087	0.130	0.174
Field size adequate	0.017	0.013	0.009	0.004	0.0

could possibly be observed in a study of 259 patients with these relapse and treatment totals. These five tables (labelled 0 through 4 for convenience) are displayed in figure 2.1; notice that the label for each 2 × 2 table corresponds to the number in the upper left-hand corner. The sample relapse rates for each of these tables are indicated in table 2.2.

Table 2.2 illustrates the folly of assuming that sample relapse rates are the same as population relapse rates. Quite obviously, the fixed study totals (259 patients, 4 relapses and 23 patients receiving inadequate radiotherapy) severely restrict the set of possible sample relapse rates that may be observed to exactly five pairs of values.

This problem with sample relapse rates suggests that perhaps we should redirect our attention from the sample to the population. Clearly, there is a population relapse rate, r_1 say, for patients who received adequate radiotherapy, and there is a second relapse rate, r_2 say, for patients whose radiotherapy was inadequate. The hypothesis which Thomas et al. wish to investigate is whether or not these relapse rates are the same, i.e., whether $r_1 = r_2$ or $r_1 \neq r_2$. If r_1 and r_2 are identical, or very similar, then we might expect this similarity to be reflected in the observed data, to the extent that the sample size, number of relapses and treatment totals permit. On the other hand, if r_1 and r_2 are quite different, then we would expect the data to reflect this difference, again, as the numbers permit.

If we suppose that r_1 and r_2 are the same, then a natural estimate of the common relapse rate $r = r_1 = r_2$, say, is 4/259 = 0.015, the overall relapse rate in the sample. Of course, r is unlikely to be exactly 0.015; nevertheless, if our sample is representative of the population, r should be fairly close to the observed relapse rate. And if 0.015 is close to r, the relapse rate in the population of Wilms' tumor patients, then among the 23 who received inadequate radiotherapy we would expect to observe approximately $23 \times 0.015 = 0.355$ relapses in the operative site. From this perspective we realize that there is an obvious ranking for tables 0 through 4, based on the observed number of relapses among the 23 patients who received inadequate radiotherapy. If the population relapse rates are the same, i.e., $r_1 = r_2$, then table 0 is most consistent with that

Table 2.3. The probability of observing each of tables 0 through 4 if $r_1 = r_2$

	Table #					Total
	0	1	2	3	4	
Probability	0.6875	0.2715	0.0386	0.0023	0.0001	1.0000

situation, table 1 is less consistent than table 0 and so on; table 4 is clearly the least consistent of all five possible sample outcomes. To express this idea in a somewhat different way, we might say that, if $r_1 = r_2$, then we would be most inclined to expect table 0 to result from the review of patient records, and least inclined to expect table 4.

Once we have realized this fact, we have begun to develop an objective method of evaluating the degree of consistency between the data and the hypothesis of interest. Nevertheless, we still have a considerable distance to cover. For example, it should be clear that because the study patients are a sample from the population, and because sampling necessarily involves uncertainty, we cannot state conclusively that, if $r_1 = r_2$, the study records would not yield table 4. Quite obviously, what we need now is an objective measure of how likely, or unlikely, each of tables 0 through 4 would be if the relapse rates r_1 and r_2 are the same. And this is precisely what probability calculations can provide. If we assume, a priori, that $r_1 = r_2$, then it is fairly simple for a statistician to calculate the probability that in a sample of 259 patients, 23 of whom received inadequate radiotherapy and 4 of whom relapsed at the operative site, exactly 0 (or 1, 2, 3 or 4) patients who relapsed also received inadequate radiotherapy. Rather than labor over the details of these calculations, we have listed the probabilities in table 2.3 and will defer the details to chapter 3.

The entries in table 2.3 indicate that, if the population relapse rates for both groups of Wilms' tumor patients are the same, then samples of size 259 having the same relapse and treatment totals (4 and 23, respectively) are most likely to yield table 0 and least likely to result in table 4. However, each of the five tables might possibly occur in a particular sampling situation. Notice, also, that table 2 (which was reported by Thomas et al.) is a relatively unlikely outcome if the equal relapse rates assumption is correct. If we consider all tables which are no more consistent with the hypothesis $r_1 = r_2$ than table 2, i.e., tables 2, 3 and 4, then by adding the corresponding probabilities we determine that data which are no more consistent with the hypothesis $r_1 = r_2$ than those which were observed (i.e., tables 2, 3 and 4) could be expected to occur in approximately four studies out of every hundred having the same sample size (259),

Table 2.4. The two possible explanations for the observed result in the study of Thomas et al. [1]

	Explanation 1	Explanation 2
Hypothesis	The relapse rates r_1, r_2 are the same (i.e., $r_1 = r_2$)	The relapse rates r_1, r_2 are different (i.e., $r_1 \neq r_2$)
Observed result	The observed difference in the sample relapse rates is due to sampling uncertainty (chance)	The observed difference in the sample relapse rates reflects the true difference between r_1 and r_2 in the population

relapse and treatment totals (4, 23, respectively). It is this value, 0.041, the sum of the probabilities for tables 2, 3 and 4, that statisticians call the significance level (p-value) of a test of the hypothesis $r_1 = r_2$.

At this point in the argument we pause to consider possible explanations for the observed result. Each explanation is based on a specific hypothesis about the population relapse rates. Clearly, to be satisfactory, an explanation should account for the significance level of the data. As table 2.4 indicates, there are basically two explanations which account for the result observed in the sample of Thomas et al.

Clearly, we cannot determine which explanation, if either, is correct; this is beyond the realm of statistics. We cannot determine directly whether Explanation 1 is more likely since, if $r_1 \neq r_2$, we do not know how great is the actual difference in relapse rates. However, we do know that if Explanation 1 is correct, the significance level ($p = 0.04$) specifies how likely we would be to obtain a study sample which is no more consistent with the hypothesis of equal relapse rates than table 2.

Which explanation is preferable, the first or the second? As a general rule in scientific research, the simplest description or model of a situation which is, at the same time, consistent with the available evidence is preferred. On this basis, Explanation 1 usually would be selected, a priori, because it does not involve a relapse rate which depends on the radiotherapy received. However, if we adopt Explanation 1, then we are also choosing to account for the observed difference in sample relapse rates on the basis of sampling uncertainty. We have calculated that such discrepancies would occur in at most four studies out of every hundred of a similar size. Statisticians normally suggest that when the significance level associated with the simpler explanation is very small, it is justifiable to conclude that the simpler explanation is not consistent with the evidence at hand; instead, the more complicated explanation should be adopted. Therefore, since the significance level of the data of Thomas et al. with respect to the hypothesis of equal relapse rates is relatively small ($p = 0.04$), we conclude from an analysis of the study records that the weight of evidence contradicts the hypothesis that $r_1 = r_2$, i.e., Explanation 2 is preferred.

2.3. Common Features of Significance Tests

The preceding example provides a valuable introduction to the general topic of significance tests. The purpose of a significance test is to determine the degree of consistency between a specific hypothesis, represented by H say, and a set of data. In most cases, H is a simple description of some aspect of a particular population and the collected data are obtained from a sample drawn from the population. While the sample is thought to be representative of the population, it is clear that a different sample would give rise to a different set of data. There are always two fairly obvious explanations for the observed results:

I The hypothesis H is true and sampling uncertainty (chance) is the reason for the observed result.

II The hypothesis H is false and the difference between the true situation and H is the reason for the observed result.

The objective measure which is used to evaluate the degree to which the data are consistent with the hypothesis H is a probability calculation of the sampling uncertainty referred to in Explanation I. If we assume H is true, we can calculate how likely, or unlikely, the outcome observed in the sample would be. If the outcome is likely (normally, this is thought to be the case when the significance level is at least 0.05, i.e., $p \geq 0.05$), then we have meager evidence, at best, for concluding that the data are inconsistent with H. However, if the outcome is unlikely when H is true (normally, this is thought to be the case when the significance level is less than 0.05, i.e., $p < 0.05$), then both I and II become plausible explanations. Provided the sampling procedure has been carefully designed to guard against gross defects, as the significance level (p-value) decreases we become less inclined to favor I, preferring II as the more plausible explanation for the apparent inconsistency between the data and the specific hypothesis H.

In summary, a significance test assumes that a specific hypothesis, H, about a population is true and compares the outcome observed in the sample with all other possible outcomes which sampling uncertainty might have generated. To carry out a test of significance, then, the following are necessary:

(1) A hypothesis about the population, usually referred to as H_0 (the null hypothesis). In the example discussed in §2.2, H_0 was the assumption that the population relapse rates, r_1 and r_2, were equal.

(2) Data from the population. These are obtained from a random sample and are usually summarized in the form of the observed value of a suitable test statistic. In the sample obtained by Thomas et al. [1] the test statistic was the number of patients who had received inadequate radiotherapy and who had subsequently relapsed in the operative site; its value in the sample was 2.

(3) A set of comparable events of which the outcome observed in the sample is only one possibility. In the example of §2.2, the five possible 2 × 2 tables displayed in figure 2.1 constitute the set of comparable events.

(4) The probability distribution of the test statistic (see 2), based on the assumption that the null hypothesis, H_0, is true. For the example we discussed in §2.2, this probability distribution is displayed in table 2.3.

(5) A ranking of all possible outcomes in the set of comparable events (see 3) according to their consistency with the null hypothesis H_0. In the example from Thomas et al. we ranked tables 0 through 4 as follows:

	Most consistent with H_0			Least consistent with H_0
Table #	0	1	2	3	4

This ranking was based on the probability distribution displayed in table 2.3.

(6) A calculation of the probability that sampling uncertainty (chance) would produce an outcome no more consistent with the null hypothesis, H_0, than the outcome which actually was observed (see 2). This probability is called the significance level of the data with respect to H_0. In the example of §2.2, the significance level of the data with respect to the hypothesis of equal relapse rates was 0.041, the probability of obtaining tables 2, 3 or 4.

Many tests of significance for standard situations have been devised by statisticians, and the items outlined in (1)–(6) above have been specified. Often, all that a researcher must do in a particular circumstance is to select an appropriate test, evaluate the prescribed test statistic using the collected data, calculate the significance level (p-value) of the data with respect to the null hypothesis, H_0, and draw an appropriate conclusion.

In succeeding chapters, we will describe particular tests of significance which are basic to medical statistics. In general, these chapters adopt a more relaxed presentation of specific significance tests than we have undertaken in this chapter. The particular test which we discussed in §2.2 is known as Fisher's test for 2 × 2 contingency tables. This test is sufficiently useful to justify a presentation in more general terms. In addition, since the specific details of the example in §2.2 are now familiar ground, a discussion of Fisher's test should ease the transition of any hesitant reader onto the new aspects of well-known territory which we shall introduce in chapter 3.

3

Fisher's Test for 2 × 2 Contingency Tables

3.1. Introduction

A surprising amount of the data gathered in medical research is binary in nature, that is, it belongs to one of only two possible outcomes. For example, patients treated for Wilms' tumor either relapse in the operative site or they do not. For this reason, the methods which statisticians have devised for analyzing binary data frequently find a natural application in medical research. Data which are binary usually can be summarized conveniently in a 2 × 2 contingency table such as the one discussed in §2.2. In these situations, it is natural to ask whether the two binary classification schemes, represented by the rows and columns of the 2 × 2 table, are associated in the population. If we assume that no such association exists, then Fisher's test determines the degree to which this hypothesis is consistent with the sample data which are summarized in the 2 × 2 contingency table.

3.2. Details of the Test

Consider a simple experiment which has only two possible outcomes, e.g., tumor presence or absence, graft rejection or acceptance, 6-month survival or death prior to that time. For convenience, we shall use the generic labels 'success' and 'failure' to describe the two outcomes. 'Success' might be tumor absence, graft acceptance or 6-month survival; then 'failure' represents tumor presence, graft rejection or death prior to 6 months, respectively. In each case, we can associate a rate or proportion with each of the categories success and

Table 3.1. A 2 × 2 table summarizing binary data collected from two groups

	Success	Failure	Total
Group 1	T	$R_1 - T$	R_1
Group 2	$C_1 - T$	$C_2 + T - R_1$	R_2
Total	C_1	C_2	N

failure; these will be numbers between 0 and 1 which have the property that their sum is always 1, since success and failure are the only possible outcomes. We could describe each rate or proportion equally well as the probability of the corresponding outcome, whether success or failure. Then if p represents the probability of success, the probability of failure will be 1 – p since the probabilities of all possible outcomes in a population always sum to 1.

In the basic model for the population, we assume that each individual has a probability of success which can be specified by a value of p between 0 and 1. Therefore, the aim of the study which is summarized in a 2 × 2 contingency table might be to compare two population groups, 1 and 2 say, with respect to their probabilities of success. To devise a test of significance suited to this purpose we must also assume:

(a) The probabilities of success for each member of Groups 1 and 2 are p_1 and p_2, respectively; i.e., within each group, the probability of success does not vary from individual to individual. Random sampling of individuals from a well-defined population of interest makes this assumption a reasonable one to adopt.

(b) For any member of either group, the outcome which occurs (success/failure) does not influence the outcome for any other individual.

Once the data for the study have been collected (number of successes, failures in each group), they can be summarized in a 2 × 2 contingency table such as the one shown in table 3.1.

The symbols R_1 and R_2 represent the total numbers of observations from Groups 1 and 2, respectively; the letter R is used because these are row totals for the table. Likewise, C_1 and C_2 represent the column totals for the categories Success and Failure. The total number of observations in the sample is N, and clearly $N = R_1 + R_2 = C_1 + C_2$. We will discuss the choice of letters appearing in the four cells of the 2 × 2 table shortly. First, however, we need to specify the null hypothesis for Fisher's test.

As we indicated above, it is quite natural in this particular situation to compare the probabilities of success in Groups 1 and 2. Therefore, the null hypothesis for Fisher's test may be expressed concisely as the statement H_0: $p_1 = p_2$, i.e., the probability of success is the same for Groups 1 and 2.

The letter T in the upper left corner of the contingency table shown in table 3.1 represents the total number of successes observed in Group 1. The variable T is the test statistic for the observed data which we referred to in §2.3. Notice that since the numbers of individuals from Groups 1 and 2 are known (R_1 and R_2, respectively), as are the numbers of successes and failures (C_1 and C_2, respectively), the remaining three entries in the table can be obtained by subtraction from row or column totals once the value of T has been determined. In effect, when the row and column totals are known, T determines the split of the successes between Groups 1 and 2; this is the principal reason that T is chosen as the test statistic for the significance test.

To determine the set of comparable events, we must consider all the possible 2 × 2 tables with row totals R_1, R_2 and column totals C_1, C_2 which might have been obtained. These can be identified by allowing the value of T to vary, beginning with the smallest possible value it can assume; this will usually be 0. Figure 3.1 displays four of these tables, provided both R_1 and C_1 are at least three and C_1 is at most R_2. As we saw in the example which was discussed in §2.2, these tables can be conveniently labelled using the value which the test statistic, T, assumes in the table. Clearly, the last table in the list will be one having a value of T which is equal to the smaller of R_1 and C_1.

To obtain the significance level of the test, we need the probability that, for known values of R_1, R_2 and C_1, the split of successes between Groups 1 and 2 is T and $C_1 - T$, respectively. To calculate this probability, we must assume that the null hypothesis is true, i.e., that $p_1 = p_2$, and also that any particular set of T successes in Group 1 and $C_1 - T$ in Group 2 is equally likely to have occurred. Certain fairly simple mathematical arguments lead to the formula

$$\frac{\binom{R_1}{t}\binom{N-R_1}{C_1-t}}{\binom{N}{C_1}}$$

for the probability that table t, i.e., the 2 × 2 table with t successes in Group 1 and $C_1 - t$ in Group 2, would be observed if these assumptions are true. The symbols $\binom{R_1}{t}$, $\binom{N-R_1}{C_1-t}$ and $\binom{N}{C_1}$ used in this expression are called binomial coefficients. The binomial coefficient $\binom{n}{j}$ can be evaluated using the formula

$$\binom{n}{j} = \frac{n(n-1)(n-2)\ldots(n-j+2)(n-j+1)}{j(j-1)(j-2)\ldots(2)(1)}.$$

However, statistical tables or software are normally used to determine the significance level of Fisher's test. In any case, the most important aspect to re-

Table 0

	Success	Failure	Total
Group 1	0	R_1	R_1
Group 2	C_1	$R_2 - C_1$	R_2
Total	C_1	C_2	N

Table 1

Group 1	1	$R_1 - 1$	R_1
Group 2	$C_1 - 1$	$R_2 - C_1 + 1$	R_2
Total	C_1	C_2	N

Table 2

Group 1	2	$R_1 - 2$	R_1
Group 2	$C_1 - 2$	$R_2 - C_1 + 2$	R_2
Total	C_1	C_2	N

Table 3

Group 1	3	$R_1 - 3$	R_1
Group 2	$C_1 - 3$	$R_2 - C_1 + 3$	R_2
Total	C_1	C_2	N

Fig. 3.1. Four possible 2 × 2 tables having row totals R_1, R_2 and column totals C_1, C_2.

member is that the probability corresponding to each possible table, i.e., each value of the test statistic T, can be calculated.

Once the numerical values of the probability distribution of T have been determined, there is a simple ranking for all the possible tables. This is based on the value of the probability corresponding to each table. If t_1 and t_2 are two possible values of T and the probability corresponding to table t_1 is greater than the probability for table t_2, then we say that table t_1 is more consistent with the null hypothesis that $p_1 = p_2$ than table t_2. On this basis we can quickly rank all the tables, i.e., all possible values of the test statistic T.

Table 3.2. The results of an experiment comparing the anti-tumor activity of two drugs in leukemic mice

	Complete remission		Total
	yes	no	
Methyl GAG	7	3	10
6-MP	2	7	9
Total	9	10	19

To complete the test, we must calculate the significance level of the data with respect to the null hypothesis. From the original 2 × 2 table which we observed and the corresponding value of T, we can determine the position of the observed 2 × 2 table in the ranking. Then, by summing the probabilities for possible tables whose rankings are no more consistent than the observed 2 × 2 table, we obtain the significance level of the data with respect to the null hypothesis that $p_1 = p_2$.

To illustrate each aspect of Fisher's test in a specific case, we consider the following example. Two drugs, methyl GAG and 6-MP, were screened in a small experiment to determine which, if either, demonstrated greater anti-tumor activity in leukemic mice. Ten mice received methyl GAG and nine were treated with 6-MP. When the experiment was ended, seven of the nine mice which had achieved complete remission belonged to the methyl GAG group. Table 3.2 summarizes the results of the experiment as a 2 × 2 contingency table.

In this particular case, mice in complete remission represent observed successes and the null hypothesis for this set of data states that $p_1 = p_2$, i.e., the probability of complete remission is the same for both drugs. The observed value of the test statistic, T, is seven, the number of complete remissions observed in mice treated with methyl GAG. Since nine, the total number of complete remissions observed, is the largest number that might have been obtained in the methyl GAG group, there are ten possible 2 × 2 tables, corresponding to the values of T from zero through nine. These tables are displayed in figure 3.2, and the probability distribution for T is given in table 3.3. From the probability distribution for T, we can determine the ranking of the ten possible tables. In this particular case it turns out that table 5 is most consistent with the null hypothesis and table 0 is the least consistent of all 10. Notice that the observed result, table 7, is fifth in the ranking, followed, in order, by tables 2, 8, 1, 9 and 0. Therefore, since tables 7, 2, 8, 1, 9 and 0 are each no more consis-

Table 0

	Complete remission		Total
	yes	no	
Methyl GAG	0	10	10
6- MP	9	0	9
Total	9	10	19

Table 1

Complete remission		Total
yes	no	
1	9	10
8	1	9
9	10	19

Table 2

Methyl GAG	2	8	10
6-MP	7	2	9
Total	9	10	19

Table 3

3	7	10
6	3	9
9	10	19

Table 4

Methyl GAG	4	6	10
6-MP	5	4	9
Total	9	10	19

Table 5

5	5	10
4	5	9
9	10	19

Table 6

Methyl GAG	6	4	10
6-MP	3	6	9
Total	9	10	19

Table 7

7	3	10
2	7	9
9	10	10

Table 8

Methyl GAG	8	2	10
6-MP	1	8	9
Total	9	10	19

Table 9

9	1	10
0	9	9
9	10	19

Fig. 3.2. The ten possible 2 × 2 tables having row totals 10, 9 and column totals 9, 10.

Table 3.3. The probability of observing each of the tables shown in figure 3.2 if the null hypothesis $H_0:p_1 = p_2$ is true

T (table #)	Probability	T (table #)	Probability
0	0.00001	5	0.3437
1	0.0009	6	0.1910
2	0.0175	7	0.0468
3	0.1091	8	0.0044
4	0.2864	9	0.00019

tent with the null hypothesis than the observed result (table 7), the significance level of the test is obtained by summing the probabilities for these six tables. The sum, 0.0698, is therefore the significance level of the data with respect to the null hypothesis that the probability of complete remission is the same for both drugs. And since the significance level is roughly 0.07, we conclude that there is no substantial evidence in the data to contradict the null hypothesis.

On the basis of this small experiment, we would conclude that the antitumor activity of methyl GAG and 6-MP in leukemic mice is apparently comparable, although further investigation might be justified. In chapter 11 we consider the related problem of estimating rates for events of interest such as the occurrence of complete remission in this study.

3.3. Additional Examples of Fisher's Test

In previous editions, this section introduced specialized statistical tables that one could use to carry out Fisher's test. Although these tables are still appropriate, the widespread availability of modern statistical software packages that routinely calculate exact or approximate significance levels for various statistical tests selected by the package user has largely eliminated the routine use of statistical tables. Consequently, we have chosen to replace the specialized statistical tables for Fisher's test by additional examples to reinforce the key aspects that characterize the use and interpretation of this commonly occurring method of assessing whether two binary classification schemes are associated in a population of interest.

As part of an experiment to investigate the value of infusing stored, autologous bone marrow as a means of restoring marrow function, a researcher administered Myleran to 15 dogs. Nine were then randomized to the treatment group and received an infusion of bone marrow, while the remaining six dogs

Table 3.4. Treatment and survival status data for 15 dogs insulted with Myleran

Bone marrow infusion	30-day survival status		Total
	yes	no	
No (control)	1	5	6
Yes (treatment)	9	0	9
Total	10	5	15

Table 3.5. Treatment and follow-up examination status data for 73 patients with simple urinary tract infection

Treatment regime	Urine culture status at follow-up		Total
	negative	positive	
Single dose	25	8	33
Multiple dose	35	5	40
Total	60	13	73

served as a control group. The experiment was ended after 30 days and the results are presented in table 3.4.

In this context, a natural question to ask is whether the probability of 30-day survival is the same in both groups of dogs, i.e., no treatment effect. By appropriately specifying the use of Fisher's exact test (which is how it is usually identified in most statistical packages) in connection with this 2 × 2 contingency table, we learn that the corresponding significance level for these data is 0.002. The obvious conclusion is that the probability of 30-day survival is significantly higher in dogs that receive an infusion of stored, autologous bone marrow.

Backhouse and Matthews [2] describe an open randomized study concerning the efficacy of treating simple urinary tract infections with a single dose (600 mg) of enoxacin, a new antibacterial agent, compared with 200 mg of the same drug twice a day for three consecutive days. Of the 73 patients who had confirmed pretreatment bacterial urine cultures, 33 received a single dose of enoxacin and 40 were randomized to the multiple-dose treatment. The observed results at a follow-up examination are summarized in table 3.5.

The obvious hypothesis of interest is whether the probability of a negative urine culture ten days following the initiation of treatment with enoxacin is the same for both groups of patients. The significance level of Fisher's exact test for these data is 0.727, indicating that, from the statistical point of view, there is no evidence to contradict the view that both a single dose of enoxacin and the multiple-dose regime are equally effective treatments for simple urinary tract infections.

Available software may well be able to cope with the extensive calculations involved in evaluating exact significance levels in Fisher's test where the sample sizes in the two groups exceed 50. However, in larger samples it is more likely that the software package defaults to an approximate version of Fisher's test which is usually adequate. Therefore, in chapter 4 we intend to describe in detail the approximate test for contingency tables.

4

Approximate Significance Tests for Contingency Tables

4.1. Introduction

Fisher's test, which we discussed in chapter 3, evaluates the exact significance level of the null hypothesis that the probability of success is the same in two distinct groups. Ideally, the exact significance level of this test is what we would prefer to know in every situation. However, unless the associated calculations have been very carefully programmed, they may be erroneous if the sample sizes are large. In such situations, accurate approximations for calculating the significance level of the test that have been used for decades are frequently the default action in modern software packages. The approximate version of Fisher's test which we discuss in the following section is known as the χ^2 (chi-squared) test for 2 × 2 tables. Another merit of this approximate version is that it helps to elucidate the nature of the comparisons that Fisher's test involves. The same approximation also applies to generalizations of Fisher's test involving classification schemes with more than two categories and more than two outcomes. Thus, after discussing the simplest version of the approximation to Fisher's test in §4.2, we intend to introduce approximate significance tests for rectangular contingency tables in §4.3.

4.2. The χ^2 Test for 2 × 2 Tables

Suppose that a 2 × 2 table, such as the one shown in table 4.1, has row and column totals which are too large to be used easily in determining the significance level of Fisher's test. For reasons of convenience in describing the χ^2 test, we have chosen to label the entries in the 2 × 2 table O_{11}, O_{12}, O_{21} and O_{22} (see table 4.1). The symbol O_{11} represents the observed number of successes in Group 1. If we call 'success' category I and 'failure' category II, then O_{11} is the

Table 4.1. A 2 × 2 table summarizing binary data collected from two groups

	Success (I)	Failure (II)	Total
Group 1	O_{11}	O_{12}	R_1
Group 2	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

number of category I observations in Group 1. Similarly, the symbol O_{12} is the number of category II observations (failures) in Group 1; the symbols O_{21} and O_{22} represent the corresponding category I (success) and category II (failure) totals for Group 2.

The assumptions on which the χ^2 test is based are the same as those for Fisher's test. If p_1 and p_2 represent the probabilities of category I for Groups 1 and 2, respectively, then we are assuming that:

(a) within each group the probability of category I (success) does not vary from individual to individual,

(b) for any member of the population, the outcome that occurs (I or II) does not influence the outcome for any other individual.

Likewise, the purpose of the χ^2 test is the same as that of Fisher's test, namely, to determine the degree to which the observed data are consistent with the null hypothesis H_0: $p_1 = p_2$, i.e., the probability of category I in the two groups is the same.

The basis for the χ^2 test is essentially this: assume that the null hypothesis, H_0, is true and calculate the 2 × 2 table which would be expected to occur based on this assumption and the row and column totals R_1, R_2, C_1 and C_2. If the observed 2 × 2 table is similar to the expected 2 × 2 table, the significance level of the data with respect to H_0 will be fairly large, say 0.5. However, if the observed 2 × 2 table is very different from the expected 2 × 2 table, the significance level of the data with respect to H_0 will be rather small, say 0.05 or less. In both cases, the approximation to Fisher's test occurs in the calculation of the significance level. And this is precisely the point at which the calculations for Fisher's test become so formidable when R_1, R_2, C_1 and C_2 are quite large.

In order to illustrate the calculations which the χ^2 test involves, we will consider the sample 2 × 2 table shown in table 4.2. The data are taken from Storb et al. [3] and summarize the outcomes of 68 bone marrow transplants for patients with aplastic anemia. Each patient was classified according to the outcome of the graft (Rejection, Yes or No) and also according to the size of the marrow cell dose which was used in the transplant procedure. The principal

Table 4.2. Graft rejection status and marrow cell dose data for 68 aplastic anemia patients

Graft rejection	Marrow cell dose (10^8 cells/kg)		Total
	<3.0	≥3.0	
Yes	17	4	21
No	19	28	47
Total	36	32	68

Table 4.3. The 2 × 2 table of expected values corresponding to the observed data summarized in table 4.1

	Success (I)	Failure (II)	Total
Group 1	e_{11}	e_{12}	R_1
Group 2	e_{21}	e_{22}	R_2
Total	C_1	C_2	N

question which the data are intended to answer is whether the size of the marrow cell dose is associated with the marrow graft rejection rate.

To carry out the approximate test of significance, we need to calculate the values which would be expected in this particular sample if the null hypothesis, H_0, is true. This table of expected values will have the same row and column totals as the observed 2 × 2 table. In the 2 × 2 table shown in table 4.3, the four entries are represented by the symbols e_{11}, e_{12}, e_{21} and e_{22} to distinguish them from the values in the observed 2 × 2 table (cf. table 4.1). The meaning of the subscripts on these symbols should be fairly obvious. The symbol e_{11} represents the expected number of category I outcomes in Group 1, while e_{21} is the corresponding expected value for Group 2; likewise, e_{12} and e_{22} are the category II expected numbers for Groups 1 and 2, respectively.

The overall success rate in the observed 2 × 2 table is C_1/N. If the null hypothesis is true, this rate is a natural estimate of the common success rate for both Group 1 and Group 2. There are R_1 individuals in Group 1; therefore, if the null hypothesis is true, the expected number of category I outcomes (success) in Group 1 would be

$$e_{11} = R_1 \times \left(\frac{C_1}{N}\right) = \frac{R_1 \times C_1}{N}.$$

Table 4.4. The 2 × 2 table of expected values corresponding to the graft rejection data summarized in table 4.2

Graft rejection	Marrow cell dose (10^8 cells/kg)		Total
	<3.0	≥3.0	
Yes	$\frac{21\times36}{68}=11.12$	$\frac{21\times32}{68}=9.88$ $=21-11.12$	21
No	$\frac{47\times36}{68}=24.88$ $=36-11.12$	$\frac{47\times32}{68}=22.12$ $=47-24.88$	47
Total	36	32	68

Notice that this is the product of the row total for Group 1 (R_1) and the column total for category I (C_1) divided by the total number of observations (N). This makes the formula particularly easy to recall and use. The other expected values in the 2 × 2 table are calculated from similar formulae, viz.

$$e_{12}=\frac{R_1\times C_2}{N},\ e_{21}=\frac{R_2\times C_1}{N}\ \text{and}\ e_{22}=\frac{R_2\times C_2}{N}.$$

Of course, since the row and column totals for the 2 × 2 table of expected numbers are already fixed (R_1, R_2, C_1 and C_2), we know from the discussion of Fisher's test (see § 3.2) that, once we have calculated e_{11}, we can obtain the other entries in the table of expected numbers by subtraction. One possible set of formulae for obtaining e_{12}, e_{21} and e_{22} in this way is

$$e_{12} = R_1 - e_{11},\ e_{21} = C_1 - e_{11}\ \text{and}\ e_{22} = R_2 - e_{21}.$$

Although we could list other sets of formulae for calculating e_{12}, e_{21} and e_{22}, these versions will suffice. Any correct set will always produce a table of expected numbers having row and column totals R_1, R_2, C_1 and C_2. Remember that these expected values are based on the assumption that the null hypothesis, H_0, is true, i.e., the probability of category I (success) is the same for Groups 1 and 2. Table 4.4 shows the table of expected values for the data concerning bone marrow transplantation, including details of the calculations.

In §2.3 we specified that an appropriate test statistic, T, is needed to carry out a test of significance. In the case of Fisher's test, T is the number of successes (category I) observed in Group 1. The choice of T, the test statistic, lies at the heart of the χ^2 approximation to Fisher's test. One formula for T is the expression

$$T = \frac{(O_{11} - e_{11})^2}{e_{11}} + \frac{(O_{12} - e_{12})^2}{e_{12}} + \frac{(O_{21} - e_{21})^2}{e_{21}} + \frac{(O_{22} - e_{22})^2}{e_{22}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}.$$

We can see here that use of the summation symbol, Σ, which we introduced in chapter 1, greatly facilitates writing the formula for T. The use of two Σ symbols simply means that, in the expression $(O_{ij} - e_{ij})^2/e_{ij}$, i is replaced by 1 and 2 and for each of these values j is replaced by 1 and 2. Thus, each of the four terms which must be summed is generated.

A slightly lengthier formula, involving an adjustment called Yates' continuity correction, specifies that

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|O_{ij} - e_{ij}| - \tfrac{1}{2})^2}{e_{ij}}.$$

The symbol |...| is mathematical shorthand which means 'use the non-negative value of the quantity between the vertical lines'. The continuity correction is obtained by subtracting 1/2 from each of the non-negative differences; this adjustment improves the probability approximation which we will discuss later. In order to calculate the value of T for a given sample of data, we must first obtain the 2 × 2 table of expected numbers and then use O_{11}, O_{12}, O_{21}, O_{22} and e_{11}, e_{12}, e_{21} and e_{22} in the formula. However, even this calculation can be reduced. Recall that the formulae for the entries in the table of expected numbers only use R_1, R_2, C_1, C_2 and N. If we replace e_{11}, e_{12}, e_{21} and e_{22} in the formula for T by their values in terms of row and column totals, it is possible to show that there is another formula for T which always gives the same value as the lengthy formula specified above. This alternative formula for T is

$$T = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - \tfrac{1}{2}N)^2}{R_1R_2C_1C_2}.$$

Apart from the fact that this latter formula for T is more concise and involves less calculation, we notice at once that the shorter version does not require any of the values e_{11}, e_{12}, e_{21} or e_{22} from the 2 × 2 table of expected numbers. Therefore, by using the shorter formula, we avoid having to calculate these values at an intermediate step; nonetheless, the two formulae are exactly equivalent. If we use the data given in table 4.2 to calculate T, the observed value of the test statistic is

$$\frac{68(|17 \times 28 - 4 \times 19| - 34)^2}{21 \times 47 \times 36 \times 32} = 8.01.$$

Though we certainly recommend use of the concise formula for T, there is an advantage in our having introduced the longer version first. For a given set of row and column totals (R_1, R_2, C_1 and C_2), the test statistic T could take on many different values, depending on the observed values O_{11}, O_{12}, O_{21} and O_{22}. Whatever this set of possible values of T might be, there are some common features which we can deduce by examining the longer formula for T. To begin with, each of the four terms in the long version of T must be positive; therefore, T will always be positive. Secondly, if each observed value, O_{ij}, is very close to its corresponding expected number, e_{ij}, then T will be very close to 0. Thus, a small value of T suggests that the observed data are consistent with the null hypothesis, i.e., the probability of category I (success) is the same for Groups 1 and 2. Conversely, if each O_{ij} is very different from the corresponding e_{ij}, then T should be rather large. Thus, a large value of T suggests that the observed data are not consistent with the null hypothesis. And so we see that if the observed value of T, for a particular set of data, is the number t_o (for example, 8.01), then values of T which exceed t_o correspond to possible 2 × 2 tables which are less consistent with the null hypothesis, H_0, than the 2 × 2 table that actually was observed ($T = t_o$). In order to calculate the significance level of the data, we simply need to sum the probabilities associated with all possible values of T which are greater than or equal to the observed value t_o, i.e., calculate $\Pr(T \geq t_o)$. In the bone marrow transplant example this is precisely $\Pr(T \geq 8.01)$.

As we indicated earlier, the heart of the χ^2 approximation to Fisher's test lies in the choice of the test statistic, T. We have already learned that 2 × 2 tables which are no more consistent with the null hypothesis than the observed 2 × 2 table ($T = t_o$) correspond to the set described by the inequality $T \geq t_o$. To obtain the probability of this set, we require the probability distribution of T when H_0 is true. And herein lies the importance of the choice of T as a test statistic. By means of complex mathematical arguments, statisticians have proved that when H_0 is true, i.e., when the probability of category I is the same in Groups 1 and 2, the test statistic T has a distribution which is approximately the same as the probability distribution called χ_1^2 (chi-squared with one degree of freedom). Therefore, we can determine the significance level of the data (recall that this is $\Pr(T \geq t_o)$) by calculating the probability that $\chi_1^2 \geq t_o$. The shorthand notation $\Pr(\chi_1^2 \geq t_o)$ represents the probability that a random variable, whose probability distribution is χ_1^2, exceeds t_o. Of course, this means that the significance level we calculate will be approximate; however, a great deal of statistical research has been devoted to showing that the approximation is quite accurate for most cases not covered by exact calculation of the significance level for Fisher's test.

To evaluate $\Pr(\chi_1^2 \geq t_o)$ we must refer to statistical tables of the χ_1^2 distribution. In §4.3 we intend to introduce other members of the χ^2 family of probability distributions. For this reason, we have chosen to discuss the use of statistical tables for χ^2 probability distributions in the final section of this chapter. For our present purposes, it suffices to indicate that $\Pr(\chi_1^2 \geq t_o)$ can be evaluated by referring to statistical tables. Therefore, the approximate significance level of the data with respect to the null hypothesis, H_0, can be determined.

To conclude the bone marrow transplant example we need to evaluate $\Pr(\chi_1^2 \geq 8.01)$, since 8.01 is the observed value of T in this case. From tables for the χ_1^2 distribution we learn that $0.001 < \Pr(\chi_1^2 \geq 8.01) < 0.005$. Since the exact value of $\Pr(\chi_1^2 \geq 8.01)$ is still an approximate significance level, even if the approximation is quite accurate, the range 0.001–0.005 is enough to indicate to us that the data represent very strong evidence against the null hypothesis that the graft rejection rate is the same for patients receiving the two different marrow cell doses. Not only does the graft rejection rate appear to be much smaller (4/32 versus 17/36) for patients transplanted with the larger marrow dose, but on the basis of the χ^2 test we can state that this apparent difference is statistically significant, since the significance level of the test is $p < 0.005$.

Comments:

(a) Depending on experimental conditions, the observed results in a 2 × 2 table may be obtained from two essentially different experimental designs:

i. The row totals R_1, R_2 are fixed by the experimenter in advance. For example, to compare 6-month survival rates under two chemotherapies, we might select 100 patients and randomize a predetermined number, say R_1, and the remainder, $100 - R_1$ (which therefore equals R_2), to the two treatment arms. Or, to compare the probability of tumor development after insult with croton oil in male and female mice, we would randomly select a fixed number of each sex. In this latter case we might have $R_1 = R_2 = 25$.

ii. The row totals R_1, R_2 are random (not fixed in advance by the experimenter). This type of design would probably be used to compare marrow graft rejection rates among patients with and without prior transfusions, since the number of patients in these latter two categories usually cannot be fixed in advance. This design might also be used to compare the probability of tumor presence in adult mice which are dominant and recessive in some genetic characteristic.

Regardless of the experimental design, the null hypothesis on which the significance test is based remains unchanged, namely H_0: $p_1 = p_2$, i.e., the probability of category I (success) is the same for both groups.

(b) As we mentioned earlier, the distribution of the test statistic T is only approximately χ_1^2. In general, the accuracy of this approximation depends on the total numbers of observations in the rows and columns of the 2 × 2 table. A conservative rule-of-thumb for ensuring that the approximation is accurate requires that all the expected numbers, e_{ij}, exceed five. A more liberal rule allows one expected number to be as low as two. If the values in the 2 × 2 table of expected numbers are small, then Fisher's test should be used to evaluate the significance level of the data with respect to H_0.

4.3. The χ^2 Test for Rectangular Contingency Tables

In the preceding section, we considered the problem of analyzing binary data (i.e., Success/Failure, Response/No Response) collected from independent samples drawn from two populations (i.e., Control, Treatment). However, the simplicity of the 2 × 2 table is also a major disadvantage if initial tabulations of the observed data are more extensive than the binary categories Success and Failure. Clearly, reducing more detailed observations to simple binary categories results in lost information. The rectangular contingency table with, say, *a* rows and *b* columns generalizes the simple 2 × 2 table.

The basic model for the rectangular contingency table assumes that the outcome for each of N experimental units (patients, animals, lab tests, etc.) may be classified according to two schemes, A and B, say. If the classification categories for scheme A are labelled $A_1, A_2, \ldots, A_a$, then the observation for each member of a sample belongs to exactly one of these categories. For example, in a clinical trial design to compare the effects of PAS and streptomycin in the treatment of tuberculosis, each participant in the trial would be treated with one of PAS or streptomycin, or perhaps a combination of both drugs. In this case, A_1 might represent the PAS-only group, A_2 the streptomycin-only group and A_3 the combined drugs group. Similarly, if the classification categories under scheme B are labelled $B_1, B_2, \ldots, B_b$, then the outcome for each member of the sample belongs to exactly one of these categories as well. In the pulmonary tuberculosis clinical trial mentioned above, the response of each patient to treatment, as measured by the analysis of a sputum sample, would be one of positive smear, negative smear and positive culture or negative smear and negative culture. In this case, we might label the patients having a positive smear B_1, those having a negative smear but positive culture B_2, and those with a negative smear and negative culture B_3.

The results of classifying each sample unit according to both schemes, A and B, may be concisely summarized in the rectangular contingency table shown in table 4.5. In this $a \times b$ contingency table with *a* rows and *b* columns, each cell corresponds to a unique row and column combination. Therefore, each cell identifies the number of sample members which were classified as

Table 4.5. A rectangular contingency table with *a* rows and *b* columns

Classification scheme A	Classification scheme B						Total
	B_1	B_2	•	•	•	B_b	
A_1	O_{11}	O_{12}	•	•	•	O_{1b}	R_1
A_2	O_{21}	O_{22}	•	•	•	O_{2b}	R_2
•	•	•				•	•
•	•	•				•	•
•	•	•				•	•
A_a	O_{a1}	O_{a2}	•	•	•	O_{ab}	R_a
Total	C_1	C_2	•	•	•	C_b	N

Table 4.6. Sputum analysis and treatment data for 273 pulmonary tuberculosis patients

Treatment	Sputum analysis			Total
	positive smear	negative smear, positive culture	negative smear, negative culture	
PAS only	56	30	13	99
Streptomycin only	46	18	20	84
Combined drugs	37	18	35	90
Total	139	66	68	273

belonging to a unique, combined category of Scheme A *and* Scheme B. Paralleling our approach to 2 × 2 tables in §4.2, we have chosen to represent by O_{11} the number of sample members which are both A_1 and B_1 Similarly, the symbol O_{ij} denotes the total number of sample members which are both A_i and B_j. For example, the Medical Research Council [4] reported the details of a clinical trial in which 273 patients were treated for pulmonary tuberculosis with one of PAS, streptomycin or a combination of the two drugs. A 3 × 3 cross-tabulation of the results of that trial is presented in table 4.6.

In general, the null hypothesis which is tested using the χ^2 test for a rectangular contingency table is H_0: the A and B classification schemes are independent. Typically, this statement is interpreted in medical situations to mean that there is no association between the two classification schemes, A and B. For example, in the clinical trial involving the treatment of pulmonary tuberculosis with PAS, streptomycin or a combination of these two drugs, the null hypothesis of independence means that a patient's condition at the conclusion of the trial did not depend on the treatment received.

However, there is an equivalent null hypothesis which is tested in exactly the same way, and which may occasionally be more appropriate. If one of the classification schemes, say A, corresponds to sampling from *a* different groups, then H_0 can also be interpreted as specifying that the probability distribution of the B classification scheme is identical in all of the *a* groups. This is simply a generalization of the null hypothesis which we discussed in §4.2 in connection with the χ^2 test in 2 × 2 contingency tables. For example, suppose that the categories of scheme A represent three strains of mice, and the categories of B describe three types of tumors which are observed in a sample of 75 mice, 25 of each strain, all of which are tumor-bearing. Then this second interpretation of the null hypothesis specifies that the proportions of the three tumor types arising in each strain are identical.

The procedure for determining the significance level of the data in an $a \times b$ contingency table with respect to the null hypothesis, H_0, is a simple generalization of the χ^2 test for a 2 × 2 table. The steps in the procedure are the following:

(1) Compute a corresponding $a \times b$ table of expected numbers based on the row totals $R_1, R_2, ..., R_a$ and the column totals $C_1, C_2, \ldots, C_b$ in the observed table. The formula for the expected number, e_{ij}, in the cell specifying the joint category A_i and B_j is

$$e_{ij} = R_i \times \left(\frac{C_j}{N}\right) = R_i \times C_j/N.$$

Notice that this number is obtained by first multiplying the unique row and column totals corresponding to the joint category A_i and B_j. The resulting product is divided by N, the overall sample size. As we indicated above, this formula for calculating e_{ij} is a simple extension of the rule which we discussed for 2 × 2 tables in §4.2. Recall, also, that in the 2 × 2 case the table of expected values could be completed by subtraction from row and column totals after calculating e_{11}. Likewise, in the $a \times b$ case, the last entry in each row and column of the table of expected numbers could be obtained by subtracting the other calculated values in the row and column from the corresponding row or column total. This means that a total of $(a - 1) \times (b - 1)$ values of e_{ij} could be calculated using the formula given above, and the remaining $(a + b - 1)$ values can then be determined by subtraction. Table 4.7 shows the table of expected values corresponding to the $a \times b$ contingency table displayed in table 4.5. Detailed calculations for the PAS, streptomycin clinical trial example, showing the 3 × 3 table of expected numbers, are given in table 4.8.

Table 4.7. The $a \times b$ table of expected values corresponding to the observed data summarized in table 4.5

Classification scheme A	Classification scheme B						Total
	B_1	B_2	•	•	•	B_b	
A_1	e_{11}	e_{12}	•	•	•	e_{1b}	R_1
A_2	e_{21}	e_{22}	•	•	•	e_{2b}	R_2
•	•	•				•	•
•	•	•				•	•
•	•	•				•	•
A_a	e_{a1}	e_{a2}	•	•	•	e_{ab}	R_a
Total	C_1	C_2	•	•	•	C_b	N

Table 4.8. The 3 × 3 table of expected values corresponding to the sputum analysis data summarized in table 4.6

Treatment	Sputum analysis			Total
	positive smear	negative smear, positive culture	negative smear negative culture	
PAS only	$\frac{99\times139}{273}=50.41$	$\frac{99\times66}{273}=23.93$	$\frac{99\times68}{273}=24.66$	99
Streptomycin only	$\frac{84\times139}{273}=42.77$	$\frac{84\times66}{273}=20.31$	$\frac{84\times68}{273}=20.92$	84
Combined drugs	$\frac{90\times139}{273}=45.82$	$\frac{90\times66}{273}=21.76$	$\frac{90\times68}{273}=22.42$	90
Total	139	66	68	273

(2) If a certain row or column in the table of expected numbers contains entries which are generally smaller than 5, it is usually wise to combine this row or column with another suitable row or column category to ensure that all of the e_{ij}'s are at least 5. Small expected numbers can seriously affect the accuracy of the approximation which is used to calculate the significance level of the data with respect to the null hypothesis. Of course, the combined categories must represent a reasonable method of classification in the context of the problem.

Treatment	Sputum analysis		
	positive smear	negative smear, positive culture	negative smear, negative culture
PAS only	$\frac{(56-50.41)^2}{50.41} = 0.62$	$\frac{(30-23.93)^2}{23.93} = 1.54$	$\frac{(13-24.66)^2}{24.66} = 5.51$
Streptomycin only	$\frac{(46-42.77)^2}{42.77} = 0.24$	$\frac{(18-20.31)^2}{20.31} = 0.26$	$\frac{(20-20.92)^2}{20.92} = 0.04$
Combined drugs	$\frac{(37-45.82)^2}{45.82} = 1.70$	$\frac{(18-21.76)^2}{21.76} = 0.65$	$\frac{(35-22.42)^2}{22.42} = 7.06$

$t_o = 0.62 + 1.54 + 5.51 + 0.24 + 0.26 + 0.04 + 1.70 + 0.65 + 7.06 = 17.62.$

Fig. 4.1. Calculating the quantities $(O_{ij} - e_{ij})^2/e_{ij}$ and t_o, the observed value of the test statistic, for the sputum analysis data.

(3) Using the two $a \times b$ tables of observed and expected numbers, calculate for each of the $a \times b$ joint categories the quantity

$$\frac{(O_{ij} - e_{ij})^2}{e_{ij}}.$$

Figure 4.1 shows a 3 × 3 table displaying these values for the PAS, streptomycin clinical trial.

(4) The test statistic, T, is the sum of the individual values of $(O_{ij} - e_{ij})^2/e_{ij}$ for all $a \times b$ joint categories. If we call the resulting sum t_o, then t_o is the observed value of the test statistic. A mathematical expression which describes all this calculation is the simple formula

$$T = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(O_{ij} - e_{ij})^2}{e_{ij}}.$$

As we explained in §4.2, the symbol $\sum_{i=1}^{a}\sum_{j=1}^{b}$ means that the $a \times b$ individual values of $(O_{ij} - e_{ij})^2/e_{ij}$ must be added together. Notice that this test statistic is similar to the one specified in §4.2 for the 2 × 2 table. However, the formula has been generalized to accommodate *a* rows and *b* columns. Notice, also, that with the more general $a \times b$ table a continuity correction is not used.

(5) As in the case of the 2 × 2 table, if t_o is the value of the test statistic for the observed data, then values of T which equal or exceed t_o, i.e., $T \geq t_o$, correspond to $a \times b$ contingency tables which are no more consistent with the null hypothesis, H_o, than the observed data. Therefore, the significance level of the data with respect to H_o is obtained by calculating $\Pr(T \geq t_o)$. Statisticians have determined that if H_o is true, i.e., if the A and B classification schemes are independent, then the probability distribution of T is approximately the same as that of a χ^2_k (chi-squared with k degrees of freedom) random variable. The value of k depends on the number of rows and columns in the rectangular contingency table; more specifically, $k = (a - 1) \times (b - 1)$. By referring to the statistical tables for the χ^2 family of distributions or otherwise, we can calculate the probability that $\chi^2_k \geq t_o$; for details regarding this calculation, see §4.4. Therefore, the significance level of the data with respect to the null hypothesis is approximately $\Pr(\chi^2_k \geq t_o)$, where t_o represents the value of the test statistic, T, for the observed data.

Figure 4.1 not only displays the nine individual values of the quantity $(O_{ij} - e_{ij})^2/e_{ij}$ but also gives the observed value of T, namely $t_o = 17.62$. Since the contingency table from which t_o was derived consisted of three rows ($a = 3$) and three columns ($b = 3$), the distribution of T, if H_0 is true, is χ^2_4. Therefore, the significance level of the data is given by $\Pr(\chi^2_4 \geq 17.62)$. Statistical tables for the χ^2 distribution show that this probability is less than 0.005. Therefore, we conclude that the data contradict the null hypothesis of independence, i.e., patient condition at the conclusion of the clinical trial seemingly is associated with the type of treatment received. The combined drugs are more efficacious in treating pulmonary tuberculosis than either drug used alone.

Notice that the expected numbers in the 3 × 3 table presented in table 4.8 were all larger than 5. In this case it was not necessary to combine rows or columns in order to improve the accuracy of the approximate significance test. However, the following example will illustrate more precisely what is involved when too many of the expected numbers are too small.

In a prospective study of venocclusive disease (VOD) of the liver, McDonald et al. [5] reviewed the records of 255 patients undergoing bone marrow transplant for malignancy. A primary purpose of the study was to investigate the relationship between pretransplant liver disease and the incidence of VOD. The patients were divided into five diagnosis groups: D_1 (primarily acute lymphocytic leukemia), D_2 (primarily acute myelogenous leukemia), D_3 (chronic myelogenous leukemia), D_4 (solid tumors) and D_5 (lymphoma, lymphosarcoma and Hodgkin's disease). Preliminary analysis of the data revealed that both pretransplant SGOT (aspartate aminotransferase) and diagnosis group were associated with the incidence of VOD. In order to determine whether SGOT level and diagnosis were associated, the cross-tabulation pre-

Table 4.9. SGOT levels and diagnosis data for 255 patients transplanted for malignancy: **a** initial tabulation; **b** revised tabulation

a Initial tabulation

SGOT levels	Diagnosis group					Total
	D_1	D_2	D_3	D_4	D_5	
Normal	70 (64.45)[1]	70 (72.91)	14 (15.62)	9 (9.11)	3 (3.91)	166
1–2 × Normal	14 (18.25)	21 (20.64)	7 (4.42)	3 (2.58)	2 (1.11)	47
2–3 × Normal	3 (5.44)	8 (6.15)	2 (1.32)	1 (0.77)	0 (0.32)	14
3–4 × Normal	3 (3.88)	6 (4.39)	1 (0.94)	0 (0.55)	0 (0.24)	10
>4 × Normal	9 (6.98)	7 (7.91)	0 (1.70)	1 (0.99)	1 (0.42)	18
Total	99	112	24	14	6	255

b Revised tabulation

SGOT levels	Diagnosis group				Total
	D_1	D_2	D_3	D_4 and D_5	
Normal	70 (64.45)[1]	70 (72.91)	14 (15.62)	12 (13.02)	166
1–2 × Normal	14 (18.25)	21 (20.64)	7 (4.42)	5 (3.69)	47
>2 × Normal	15 (16.30)	21 (18.45)	3 (3.96)	3 (3.29)	42
Total	99	112	24	20	255

[1] The values in parentheses are expected numbers if the row and column classifications are independent.

sented in table 4.9a was prepared. If SGOT levels and diagnosis group are independent, the expected values which are given in parentheses involve too many small values to carry out a reliable test of the null hypothesis of independence. However, combining the diagnosis groups D_4 and D_5, and also the SGOT categories 2–3 ×, 3–4 × and >4 × Normal results in a 3 × 4 table with larger expected numbers in many of the cells (see table 4.9b). Notice, too, that in general, the observed and expected numbers agree fairly closely in the cells which disappear through the combining of rows and columns. The observed value of the test statistic, T, was calculated for the revised table and its value was determined to be $t_o = 4.52$. Since the revised table has three rows ($a = 3$) and four columns ($b = 4$), the approximate χ^2 distribution of T has $2 \times 3 = 6$ degrees of freedom. Therefore, the significance level of the data with respect to the null hypothesis of independence is $\Pr(\chi^2_6 \geq 4.52)$, which exceeds 0.25. The final conclusion of this analysis – which could be better carried out using methods that we will discuss in chapter 11 – is that, with respect to VOD, pretransplant SGOT levels and patient diagnosis are not related.

4.4. Using Statistical Tables of the χ^2 Probability Distribution

In §1.2 we introduced the idea that, if X is a continuous random variable, probabilities for X are represented by the area under the corresponding probability curve and above an interval of values for X. The χ^2 family of probability distributions is a set of continuous random variables with similar, but distinct, probability curves. Each member of the family is uniquely identified by the value of a positive parameter, k, which is known, historically, as the 'degrees of freedom' of the distribution. This indexing parameter, k, is usually attached to the χ^2 symbol as a subscript. For example, χ^2_4 identifies the χ^2 probability distribution which has four degrees of freedom. A sketch of the χ^2_4, χ^2_7 and χ^2_{10} probability curves is given in figure 4.2a. Although the indexing parameter k is always called degrees of freedom, it can be shown that the mean of χ^2_k is exactly k and the variance is 2k. Therefore, we could choose to identify χ^2_4, for example, as the χ^2 probability distribution with mean four. However, convention dictates that it should be called the χ^2 distribution with four degrees of freedom.

Thus far, the only reason we have required the χ^2 probability distribution has been the calculation of significance levels for the approximate test in 2 × 2 or larger contingency tables. In §4.2 we learned that if t_o is the observed value of the test statistic T, then the significance level of the corresponding null hypothesis is approximately $\Pr(\chi^2_1 \geq t_o)$. Similarly, in §4.3 we discovered that if t_o is the observed value of the test statistic computed for an $a \times b$ contingency table, then the significance level of the corresponding null hypothesis is approximately $\Pr(\chi^2_k \geq t_o)$, where $k = (a - 1) \times (b - 1)$. In principle, the values of these observed significance levels can probably be obtained from available statistical software. Nonetheless, for reasons of historical continuity, and to reinforce the concepts that underly all tests of significance, we believe that introducing how statistical tables of the χ^2 probability distribution are used is a worthwhile exercise. Thus, we require tables for each possible χ^2 probability distribution, i.e., each value of k, the degrees of freedom. These tables should specify the area under the χ^2_k probability curve which corresponds to the interval $\chi^2_k \geq t_o$; such intervals are known as right-hand tail probabilities. And since we cannot predict values of t_o in advance, we seemingly require a table for every possible value of t_o. Since this would require a separate statistical

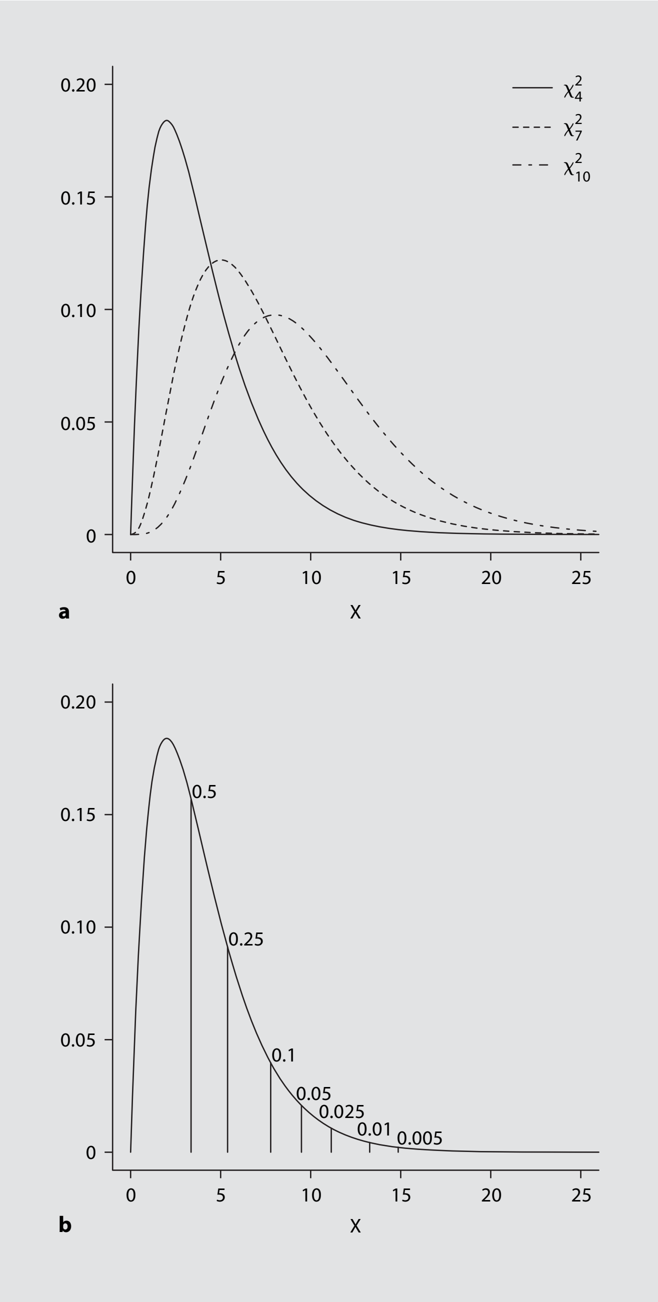


Fig. 4.2. Chi-squared probability curves. **a** χ^2_4, χ^2_7 and χ^2_{10}. **b** The location of selected critical values for χ^2_4.

Table 4.10. Critical values of the χ^2 distribution; the table gives values of the number t_o such that $\Pr(\chi^2_k \geq t_o) = p$

Degrees of freedom (k)	Probability level, p						
	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	6.626	9.236	11.07	12.83	15.09	16.75	20.52
6	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70

Abridged from Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol 1, pp 136–137. It appears here with the kind permission of the publishers.

table for every degree of freedom, it would appear that χ^2 tables must be very unwieldy.

Although sets of χ^2 tables do exist which follow the pattern we have described above, a much more compact version has been devised, and these are usually quite adequate. Since the significance level of the χ^2 test is approximate, it is frequently satisfactory to know a narrow range within which the approximate significance level of a test lies. It follows that, for any member of the χ^2 family of probability distributions, we only require a dozen or so special reference points, called critical values. These are values of the random variable which correspond to specified significance levels or, equivalently, values which cut off predetermined amounts of probability (area) in the right-hand tail of the probability curve. For example, we might wish to know the critical values for χ^2_4 which correspond to the right-hand tail probabilities 0.5, 0.25, 0.10, 0.05, 0.025, 0.01 and 0.005. The actual numbers which correspond to these probability levels for χ^2_4 are 3.357, 5.385, 7.779, 9.488, 11.14, 13.28 and 14.86. Figure 4.2b shows a sketch of the χ^2_4 probability curve indicating the location of these critical values and the corresponding right-hand tail probabilities. Such a table requires only one line of critical values for each degree of freedom. Thus, an entire set of tables for the χ^2 family of probability distributions can be presented on a single page. This is the method of presentation which most χ^2 tables adopt, and an example of χ^2 statistical tables which follow this format is reproduced in table 4.10. The rows of the table correspond to different values of k, the degrees of freedom parameter. The columns identify predetermined right-hand tail probabilities, i.e., 0.25, 0.05, 0.01, etc., and the entries in the table specify the corresponding critical values for the χ^2 probability distribution identified by its unique degrees of freedom.

The actual use of χ^2 tables is very straightforward. In order to calculate $\Pr(\chi^2_k \geq t_o)$, locate the row for k degrees of freedom. In that row, find the two numbers, say t_L and t_U, which are closest to t_o so that $t_L \leq t_o \leq t_U$. For example, if we want to evaluate $\Pr(\chi^2_4 \geq 10.4)$, these numbers are $t_L = 9.488$ and $t_U = 11.14$ since $9.488 < 10.4 < 11.14$. Since $t_L = 9.488$ corresponds to a right-hand tail probability of 0.05 and $t_U = 11.14$ corresponds to a right-hand tail probability of 0.025, it follows that $\Pr(\chi^2_4 \geq 10.4)$ is smaller than 0.05 and larger than 0.025, i.e., $0.05 > \Pr(\chi^2_4 \geq 10.4) > 0.025$. And if $\Pr(\chi^2_4 \geq 10.4)$ is the approximate significance level of a test, then we know that the p-value of the test is between 0.025 and 0.05.

5

Some Warnings Concerning 2 × 2 Tables

5.1. Introduction

In the two preceding chapters, we discussed a significance test for 2 × 2 tables such as the one shown in table 5.1. Although this version of a 2 × 2 table contains different symbols for the row totals, column totals and cell entries, the change in notation considerably simplifies the presentation of the two topics addressed in this chapter.

As the title suggests, the use of either Fisher's test or the χ^2 test to analyze a set of data is so straightforward that researchers are tempted to use one test or the other even when the use of either test is not appropriate. As a cautionary note then, in this chapter we intend to identify two situations which appear tailor-made for 2 × 2 tables but which, in fact, properly require a modified analysis. The first situation concerns the issue of combining 2×2 tables, while the second problem involves the proper method for analyzing paired binary data. In each case, the principles which are involved, namely stratification and pairing, are important statistical concepts which also apply to many other methods for analyzing data.

5.2. Combining 2 × 2 Tables

Both Fisher's test (chapter 3) and the χ^2 test outlined in chapter 4 determine the degree to which the data summarized in a 2 × 2 table are consistent with the null hypothesis that the probability of success is the same in two distinct groups. The two important assumptions of either test which we noted previously are that the probability of success must be the same for each subject

Table 5.1. A 2 × 2 table summarizing binary data collected from two groups

	Success	Failure	Total
Group 1	a	A – a	A
Group 2	b	B – b	B
Total	r	N – r	N

in a group, and that the outcome of the experiment for any subject may not influence the outcome for any member of either group.

Unfortunately, the first of these assumptions, namely that the probability of success is the same for each member of a group, is often not true; there may be factors other than group membership which influence the probability of success. If the effect of such factors is ignored and the observed data are summarized in a single 2 × 2 table, the test of significance that is used to analyze the data could mask important effects or generate spurious ones. The following example, which is similar to one described in Bishop [6], illustrates this phenomenon.

A study of the effect of prenatal care on fetal mortality was undertaken in two different clinics; the results of the study are summarized, by clinic, in figure 5.1a. Clearly, there is no evidence in the data from either clinic to suggest that the probability or rate of fetal mortality varies with the amount of prenatal care delivered. However, if we ignore the fact that this rate is roughly three times higher in Clinic 2 and combine the data in the single 2 × 2 table shown in figure 5.1b, the combined data support the opposite conclusion. Why? Notice that in Clinic 1 there are fewer deaths overall and much more intensive care, while in Clinic 2 there are more deaths and much less intensive care. Therefore, the significant difference in the two rates of fetal mortality observed in the combined table is due to the very large number of more intensive care patients contributed by Clinic 1 with its low death rate.

The preceding example illustrates only one of the possible hazards of combining 2 × 2 tables. If there are other factors, in addition to the characteristic of major interest, which might affect the probability of success, then it is important to adjust for these other factors, which are sometimes called confounding factors, in analyzing the data. In the preceding example, the primary focus of the study is the relationship between the amount of prenatal care and the rate of fetal mortality. However, we observed that the amount of prenatal care received depended on the additional factor representing clinic location (Clinic 1 or Clinic 2) which also influenced the rate of fetal mortality. Thus, clinic

a

Prenatal care	Clinic 1		Clinic 2	
	L	M	L	M
Died	12	16	34	4
Survived	176	293	197	23
Total	188	309	231	27
Fetal mortality rate	0.064	0.055	0.173	0.174
Observed value of T for χ^2 test	$t_o = 0.13$		$t_o = 0.09$	
Significance level of χ^2 test	> 0.50		> 0.75	

b

Prenatal care	L	M
Died	46	20
Survived	373	316
Total	419	336
Fetal mortality rate	0.11	0.06
Observed value of T for χ^2 test	$t_o = 5.29$	
Significance level of χ^2 test	< 0.025	

Fig. 5.1. Details of the analysis of a study of fetal mortality and the amount of prenatal health care delivered (L ≡ less, M ≡ more). **a** By individual clinic. **b** Combining over clinics.

location is a confounding factor. To properly assess the relationship between amount of prenatal care and fetal mortality, it was necessary to separately consider the data for each clinic. This simple technique of separately investigating the primary question for different cases of a confounding factor is known as stratifying the data; thus, to adjust the analysis of the study for the possible confusion which clinic location would have introduced, the study data were stratified by clinic location, i.e., divided into the data from Clinic 1 and the data from Clinic 2. Whenever it is thought necessary to adjust the analysis of a 2 × 2 table for possible confounding factors, the simplest way to effect this adjustment is to stratify the study data according to the different cases of the possible confounding factor or factors. Suppose, for the sake of illustration, that there are a total of k distinct cases (statisticians often call these cases ‘lev-

Table 5.2. A 2 × 2 table summarizing the binary data for level i of a confounding factor

	Confounding factor level i		Total
	success	failure	
Group 1	a_i	$A_i - a_i$	A_i
Group 2	b_i	$B_i - b_i$	B_i
Total	r_i	$N_i - r_i$	N_i

els') for a possible confounding factor, e.g., k different clinics participating in the fetal mortality and prenatal care study. Stratifying the study data on the basis of this confounding factor will give us k distinct 2 × 2 tables like the one shown in table 5.2, where the possible values of i, representing the distinct levels of the confounding factor, could be taken to be the numbers 1 through k.

As we saw in chapter 4, the χ^2 test for a single 2 × 2 table evaluates the discrepancy between the observed and expected values for each cell in the table, assuming that the row and column totals are fixed and the probability of success in the two groups is the same. The test of significance which correctly combines the results in all k stratified 2 × 2 tables calculates, for each distinct 2 × 2 table, an expected value, e_i, corresponding to a_i; this expected value, e_i, is based on the usual assumptions that the row and column totals A_i, B_i, r_i and $N_i - r_i$ are fixed and that the probability of success in the two groups is the same. Notice, however, that the probability of success may now vary from one stratum (2 × 2 table) to another; the test no longer requires that the probability of success must be the same for each level of the confounding factor. Instead, the null hypothesis for this test of significance specifies that, for each distinct level of the confounding factor, the probabilities of success for Groups 1 and 2 must be the same; however, this hypothesis does allow the common probabilities of success to differ from one level of the confounding factor to another. In terms of the fetal mortality example, the null hypothesis requires a constant fetal mortality rate in Clinic 1, regardless of prenatal care received, and a constant, but possibly different, fetal mortality rate in Clinic 2, regardless of prenatal care received.

Of course, evaluating the expected numbers $e_1, e_2, \ldots, e_k$ automatically determines the corresponding expected values for the remaining three cells in each of the k distinct 2 × 2 tables. As we might anticipate, the test statistic evaluates the discrepancy between the sum of the k observed values, $O = a_1 + a_2 + \ldots + a_k = \sum_{i=1}^{k} a_i$, and the sum of the corresponding k expected values, E =

2 × 2 Table

Prenatal care	Clinic 1			Clinic 2		
	L	M	Total	L	M	Total
Died	$12 = a_1$	16	$28 = A_1$	$34 = a_2$	4	$38 = A_2$
Survived	176	293	$469 = B_1$	197	23	$220 = B_2$
Total	$188 = r_1$	309	$497 = N_1$	$231 = r_2$	27	$258 = N_2$

	Clinic 1	Clinic 2
Observed value	$a_1 = 12$	$a_2 = 34$
Expected value	$e_1 = \frac{188 \times 28}{497} = 10.59$	$e_2 = \frac{231 \times 38}{258} = 34.02$
Variance	$V_1 = \frac{188 \times 309 \times 28 \times 469}{497^2 \times 496} = 6.23$	$V_2 = \frac{231 \times 27 \times 38 \times 220}{258^2 \times 257} = 3.05$

$O = 12 + 34 = 46,\quad E = 10.59 + 34.02 = 44.61,\quad V = 6.23 + 3.05 = 9.28$

$$t_o = \frac{(|46 - 44.61| - \tfrac{1}{2})^2}{9.28} = 0.09$$

Fig. 5.2. Details of the correct analysis of the fetal mortality data, adjusting for the confounding effect of clinic location.

$e_1 + e_2 + \ldots + e_k = \sum_{i=1}^{k} e_i$. To determine the significance level of the test we need to compute:

(1) $O = \sum_{i=1}^{k} a_i, E = \sum_{i=1}^{k} e_i$ where $e_i = r_i A_i / N_i$;

(2) $V = V_1 + V_2 + \ldots + V_k = \sum_{i=1}^{k} V_i$ where $V_i = \frac{r_i(N_i - r_i)A_i B_i}{N_i^2(N_i - 1)}$;

(3) the observed value, say t_o, of the test statistic $T = \frac{(|O - E| - \tfrac{1}{2})^2}{V}$.

The details of these calculations for the fetal mortality study we have been discussing are given in figure 5.2. For this set of data the observed value of the test statistic, T, is $t_o = 0.09$.

If the null hypothesis is true, the test statistic, T, has a distribution which is approximately χ_1^2. Therefore, the significance level of the test, which is $\Pr(T \geq t_o)$, can be determined by evaluating the probability that $\chi_1^2 \geq t_o$, i.e.,

Table 5.3. The results of a study to determine the diagnostic consistency between two pathologists: (**a**) initial tabulation; (**b**) revised presentation

a Initial tabulation

	Malignant	Benign	Total
Pathologist A	18	82	100
Pathologist B	10	90	100
Total	28	172	200

b Revised presentation

Pathologist B	Pathologist A		Total
	malignant	benign	
Malignant	9	1	10
Benign	9	81	90
Total	18	82	100

$\Pr(\chi_1^2 \geq t_o)$. In the case of the fetal mortality study, the approximate significance level is $\Pr(\chi_1^2 \geq 0.09) > 0.25$. Therefore, after adjusting the analysis for the confounding effect of clinic location, we conclude that there is no evidence in the data to suggest that the rate of fetal mortality is associated with the amount of prenatal care received.

Much more has been written about the hazards of combining 2 × 2 tables. For example, an exact test of the null hypothesis we have just discussed can be performed. However, the details of that version of the test are beyond the intent and scope of this brief discussion. In our view, it suffices to alert the reader to the hazards involved. For situations more complicated than the one which we have outlined, we suggest consulting a statistician.

5.3. Matched Pairs Binary Data

A second situation which, at first glance, seems tailored to the straightforward use of Fisher's test or the χ^2 test of significance in a 2 × 2 table is that of binary data which incorporate matching. The following example illustrates more precisely the situation we have in mind.

Two pathologists each examine coded material from the same 100 tumors and classify the material as malignant or benign. The investigator conducting

Table 5.4. A 2 × 2 table indicating the conclusion of each pathologist concerning tumor sample i

	Malignant	Benign	Total
Pathologist A	a_i	$A_i - a_i$	$A_i = 1$
Pathologist B	b_i	$B_i - b_i$	$B_i = 1$
Total	r_i	$N_i - r_i$	$N_i = 2$

the study is interested in determining the extent to which the pathologists differ in their assessments of the study material. The results could be recorded in the 2 × 2 table shown in table 5.3a.

Although the data are presented in the form of a 2 × 2 table, certain facets of the study have been obscured. The total number of tumors involved appears to be 200, but in fact there were only 100. Also, some tumors will be more clearly malignant than the rest; therefore, the assumption that there is a constant probability of malignancy being coded for each tumor is unreasonable. Finally, the table omits important information about the tumors which A and B classified in the same way and those on which they differed.

A better presentation of the study results might be the 2 × 2 table shown in table 5.3b, but this is still not a 2 × 2 table to which we could properly apply either Fisher's test or the χ^2 test of significance discussed in chapters 3 and 4. Though the observations are independent (there are exactly 100 tumors, each classified by A and by B), it is still unreasonable to suppose that for each of B's malignant tumors, and separately for each of B's benign tumors, there is a constant probability that A will identify the same material as malignant. Nevertheless, this is one of the two principal assumptions of both the χ^2 test and Fisher's test. (Recall that, within each group, the probability of success must be constant. In this example, the two groups are B's malignant and benign tumors.)

An appropriate way to present and analyze these data is as a series of 2 × 2 tables, with each table recording the experimental results for one tumor. A sample 2 × 2 table is shown in table 5.4. It should be immediately apparent that the method for analyzing this series of 2 × 2 tables is the procedure described in the preceding section. Because the study material is so variable, each sample item represents a distinct case of the confounding factor 'tumor material'. Therefore, we need to analyze the study by adjusting the analysis for the confounding effect of tumor material. If we do this, the 90 2 × 2 tables which correspond to tumors on which the pathologists were agreed contribute nothing to the numerator, $(|O - E| - ½)^2$, and denominator, V, of the test statistic

$$T = \frac{(|O - E| - ½)^2}{V}.$$

This occurs because whenever the pathologists agree, either $a_i = 1$, $e_i = 1$, $r_i = 2$ and $N_i - r_i = 2 - 2 = 0$ or $a_i = 0$, $e_i = 0$, $r_i = 0$ and $N_i - r_i = 2 - 0 = 2$, i.e., the net contribution to O – E is either 1 – 1 = 0 or 0 – 0 = 0 and $V_i = 0$ in both cases since one of r_i, $N_i - r_i$ is always zero. Thus, only information from 'discordant pairs' contributes to a test of the null hypothesis that, for each tumor, the probability that the specimen is labelled malignant is the same for the two pathologists. We will refer to this null hypothesis as diagnostic consistency between the pathologists. A moment's reflection will verify that if the pathologists' diagnoses were always the same, there would be no statistical evidence to contradict the null hypothesis that they are equally inclined to diagnose a tumor as malignant. Therefore, it makes sense that only tumors on which their diagnoses are different should provide possible evidence to the contrary.

As a final, cautionary note we add that care should be exercised in using the test statistic, T, with matched pairs binary data. A rough rule-of-thumb specifies that there should be at least ten disagreements or discordant pairs. For situations involving fewer than ten, there is a fairly simple calculation which yields the exact significance level of the test, and we suggest consulting a statistician in these circumstances. For the example we have been discussing, the exact significance level of a test of the null hypothesis that there is diagnostic consistency between the pathologists is 0.0215. On the other hand, if we use the test statistic, T, it turns out that O = 9, E = 5 and V = 2.5; therefore, the observed value of T is

$$t_o = \frac{(|9 - 5| - ½)^2}{2.5} = 4.90.$$

According to table 4.10, the 0.05 and 0.025 critical values for the χ_1^2 probability distribution are 3.84 and 5.02, respectively. Therefore, we know that the approximate significance level of the test is between 0.025 and 0.05. This compares favorably with the exact value of 0.0215 which we quoted previously, and points to the conclusion that the data represent moderate evidence against the null hypothesis of diagnostic consistency. When the diagnoses of the two pathologists disagree, pathologist A is much more inclined to classify the tumor material as malignant than is pathologist B.

Though we have not, by any means, exhausted the subject of analyzing binary data, at the same time not all data are binary in nature. While interested readers may wish to divert their attention to more advanced treatments of this subject, we continue our exposition of statistical methods used in medical research by discussing, in chapter 6, the presentation and analysis of survival data.

6

Kaplan-Meier or 'Actuarial' Survival Curves

6.1. Introduction

In medical research, it is often useful to display a summary of the survival experience of a group of patients. We can do this conceptually by considering the specified group of patients as a random sample from a much larger population of similar patients. Then the survival experience of the available patients describes, in general terms, what we might expect for any patient in the larger population.

In chapter 1, we briefly introduced the cumulative probability function. With survival data, it is convenient to use a related function called the survival function, $\Pr(T > t)$. If T is a random variable representing survival time, then the survival function, $\Pr(T > t)$, is the probability that T exceeds t units. Since the cumulative probability function is $\Pr(T \leq t)$, these two functions are related via the equation

$$\Pr(T > t) = 1 - \Pr(T \leq t).$$

If $\Pr(T > t)$ is the survival function for a specified population of patients, then, by using a random sample of survival times from that population, we would like to *estimate* the survival function. The concept of estimation, based on a random sample, is central to statistics, and other examples of estimation will be discussed in later chapters. Here we proceed with a very specific discussion of the estimation of survival functions.

A graphical presentation of a survival function is frequently the most convenient. In this form it is sometimes referred to as a survival curve. Figure 6.1 presents the estimated survival curve for 31 individuals diagnosed with lymphoma and presenting with clinical symptoms. The horizontal axis represents time since diagnosis and the vertical axis represents the probability or chance

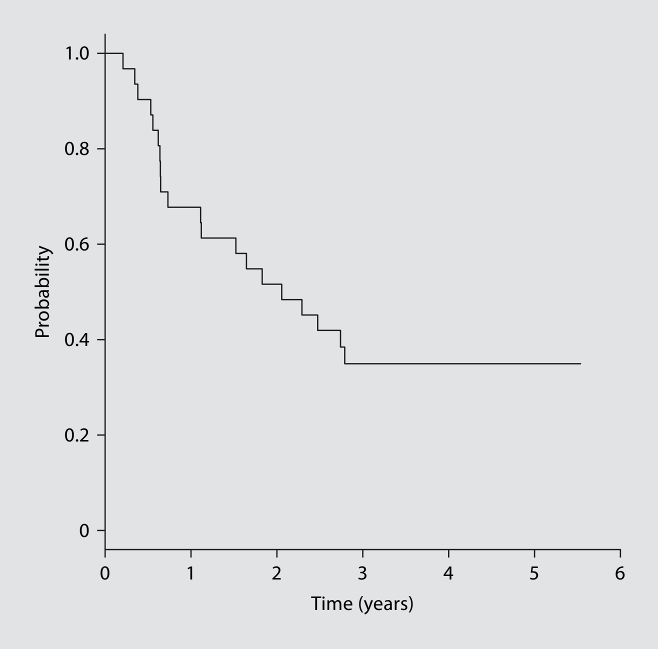


Fig. 6.1. The estimated survival curve for 31 patients diagnosed with lymphoma and presenting with clinical symptoms.

of survival. For example, based on this group of 31 patients, we would estimate that 60% of similar patients should survive at least one year, but less than 40% should survive for three years or more following diagnosis.

The estimation of survival curves like the one presented in figure 6.1 is one of the oldest methods of analyzing survival data. The early methodology is due to Berkson and Gage [7], and is also discussed by Cutler and Ederer [8]. Their method is appropriate when survival times are grouped into intervals and the number of individuals dying in each interval is recorded. This approach also allows for the possibility that individuals may be lost to follow-up in an interval. Such events give rise to censored survival times, which are different from observed survival times. Survival curves based on the methodology of Berkson and Gage are frequently referred to as 'actuarial' curves because the techniques used parallel those employed by actuaries.

The grouping of survival times may be useful for illustrative and computational purposes. However, with the increased access to computers and good statistical software which has emerged in recent years, it is now common practice to base an analysis on precise survival times rather than grouped data.

Figure 6.1 actually displays a 'Kaplan-Meier' (K-M) estimate of a survival curve. This estimate was first proposed in 1958 by Kaplan and Meier [9]. The K-M estimate is also frequently called an actuarial estimate, because it is closely related to the earlier methods. In this chapter, we will restrict ourselves to a discussion of the Kaplan-Meier estimate in order to illustrate the most important concepts. We will typically use survival time as the variable of interest, although the methodology can be used to describe time to any well-defined endpoint, for example, relapse.

6.2. General Features of the Kaplan-Meier Estimate

If we have recorded the survival times for n individuals and r of these times exceed a specified time t, then a natural estimate of the probability of surviving more than t units would be r/n. This is the estimate which would be derived from a Kaplan-Meier estimated survival curve. However, the Kaplan-Meier methodology extends this natural estimate to the situation when not all the survival times are known exactly. If an individual has only been observed for t units and death has not occurred, then we say that this individual has a censored survival time; all we know is that the individual's survival time must exceed t units. In order to illustrate the general features of the Kaplan-Meier estimate, including the methodology appropriate for censored survival times, we consider the following simple example.

Figure 6.2a presents data from a hypothetical study in which ten patients were enrolled. The observations represent the time, in days, from treatment to death. Five patients were observed to die and the remaining five have censored survival times. From these data we intend to construct an estimate of the survival curve for the study population.

Although one observation is censored at Day 1, no patients are recorded as dying prior to Day 3 following treatment. Therefore, we estimate that no deaths are likely to occur prior to Day 3 and say that the probability of surviving for at least three days is 1. As before, we use the symbol $\Pr(T > t)$ to represent the probability that T, the survival time from treatment to death, exceeds t units. Based on the study data, we would estimate that $\Pr(T > t) = 1$ for all values of t less than three days.

Nine individuals have been observed for at least three days, with one death recorded at Day 3. Therefore, the natural estimate of $\Pr(T > 3)$, the probability of surviving more than three days, is 8/9. Since no deaths were recorded between Days 3 and 5, this estimate of 8/9 will apply to $\Pr(T > t)$ for all values of t between Day 3 and Day 5 as well.

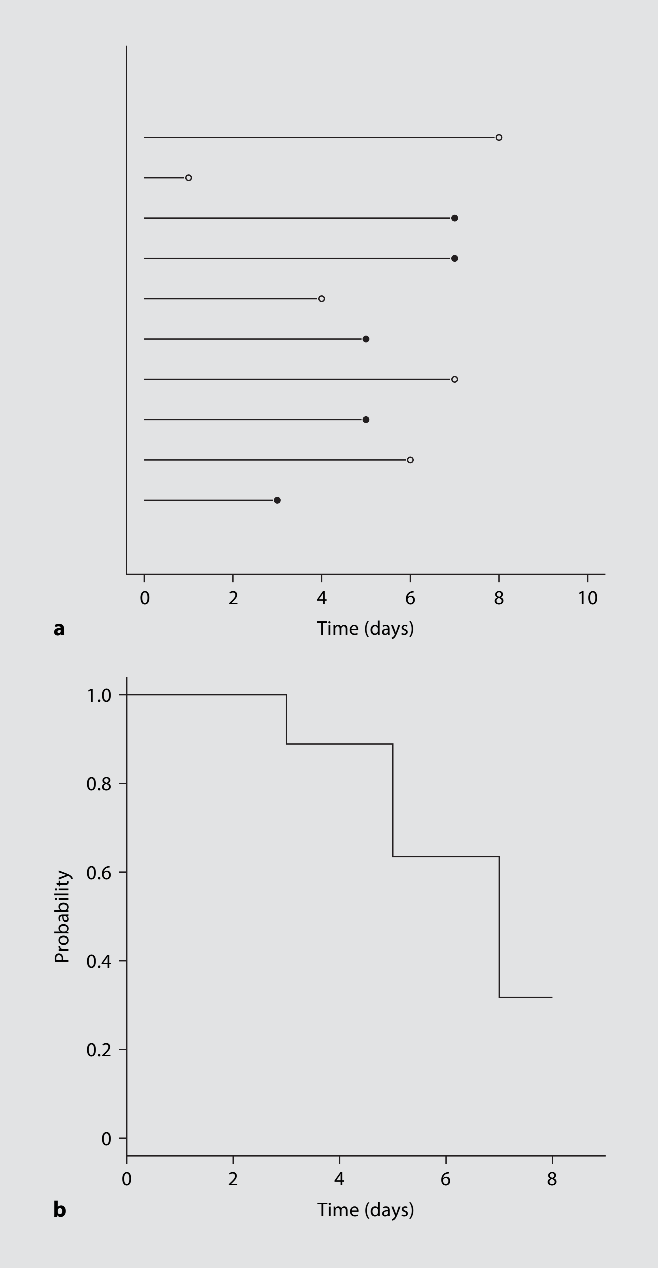


Fig. 6.2. A hypothetical study involving ten patients. **a** Survival times, in days, from treatment to death (● ≡ death, ○ ≡ censored). **b** The Kaplan-Meier estimated probability of survival function.

At Day 5 following treatment, two deaths are recorded among the seven patients who have been observed for at least five days. Therefore, among patients who survive until Day 5, the natural estimate of the probability of surviving for more than five days is 5/7. However, this is not an estimate of Pr(T > 5) for all patients, but only for those who have already survived until Day 5. The probability of survival beyond Day 5 is equal to the probability of survival until Day 5 multiplied by the probability of survival beyond Day 5 for patients who survive until Day 5. Based on the natural estimates from our hypothetical study, this product is $\frac{8}{9} \times \frac{5}{7} = \frac{40}{63}$. This multiplication of probabilities characterizes the calculation of a Kaplan-Meier estimated survival curve.

No further deaths are recorded in our example until Day 7, so that the estimate 40/63 corresponds to Pr(T > t) for all values of t between Day 5 and Day 7.

Four individuals in the study were followed until Day 7; two of these died at Day 7, one is censored at Day 7 and one is observed until Day 8. It is customary to assume that when an observed survival time and a censored survival time have the same recorded value, the censored survival time is larger than the observed survival time. Therefore, the estimate of survival beyond Day 7 for those patients alive until Day 7 would be 2/4. Since the estimate of survival until Day 7 is 40/63, the overall estimate of survival beyond Day 7 is $\frac{40}{63} \times \frac{2}{4} = \frac{20}{63}$, and so the estimate of Pr(T > t) is $\frac{20}{63}$ for all values of t exceeding 7 and during which at least one patient has been observed. The largest observation in the study is eight days; therefore Pr(T > t) = 20/63 for all values of t between Day 7 and Day 8. Since we have no information concerning survival after Day 8, we cannot estimate the survival curve beyond that point. However, if the last patient had been observed to die at Day 8, then the natural estimate of the probability of survival beyond Day 8 for individuals surviving until Day 8 would be zero (0/1). In this case, the estimated survival curve would drop to zero at Day 8 and equal zero for all values of t exceeding 8.

Figure 6.2b presents the Kaplan-Meier estimated probability of survival function for our hypothetical example. The graph of the function has horizontal sections, with vertical steps at the observed survival times. This staircase appearance may not seem very realistic, since the probability of survival function for a population is generally thought to decrease smoothly with time. Nevertheless, we have not observed any deaths in the intervals between the changes in the estimated function, so that the staircase appearance is the form most consistent with our data. If we had been able to observe more survival times, the steps in our estimated function would become smaller and smaller, and the graph would more closely resemble a smooth curve.

The staircase appearance, the drop to zero if the largest observation corresponds to a death, and the undefined nature of the estimated probability if the largest observation time is censored, may appear to be undesirable characteristics of the Kaplan-Meier estimate of the probability of survival function. All of these features arise because the methodology attempts to estimate the survival function for a population without assuming anything regarding its expected nature, and using only a finite number of observations to provide information for estimation purposes. In some sense, therefore, these undesirable characteristics are artefacts of the statistical procedure. In practical terms, these features present no serious problems since their effects are most pronounced at points in time when very few individuals have been observed. For this reason, it would be unwise to derive any important medical conclusions from the behavior of the estimated survival function at these time points. Overall, the Kaplan-Meier estimate provides a very useful summary of survival experience and deserves its pre-eminent position as a method of displaying survival data.

Comments:

(a) One rather common summary of survival experience is the sample median survival time. This statistic is probably used more widely than is warranted; nevertheless, it is a useful benchmark. Censored data can complicate the calculation of the median survival time and, as a result, a variety of estimates can be defined. A simple indication of the median survival time can be read from a Kaplan-Meier estimated survival curve as the specific time t at which $\Pr(T > t) = 0.5$. In figure 6.2b, this value may be identified as the time at which the estimated curve changes from more than 0.5 to less than 0.5. However, the estimated curve may be horizontal at the 0.5 level, in which case no unique number can be identified as the estimated median. The midpoint of the time interval over which the curve equals 0.5 is probably as reasonable an estimated median as any other choice in this situation. Use of the Kaplan-Meier estimated survival curve to estimate the median survival time ensures that correct use is made of censored observations in the calculation, and this is important.

As we noted earlier, if the largest observation has been censored, the K-M estimate can never equal zero and will be undefined when t exceeds this largest observation. In this case, if the K-M estimate always exceeds 0.5, then there can be no estimated median survival time. All that can be stated is that the median exceeds the largest observation.

(b) Another peculiar feature of the K-M estimated survival curve, especially at more distant times on the horizontal axis, is the presence of long horizontal lines, indicating no change in the estimated survival probability over a long period of time. It is very tempting to regard these flat portions as evidence of a 'cured fraction' of patients or a special group characterized in a similar way. Usually, these horizontal sections arise because only a few individuals were still under observation, and no particular importance should be ascribed to these 'long tails'. If the existence of such a special group of patients, such as a cured fraction, is thought to be likely, then it would be wise to consult a statistician concerning specialized methods for examining this hypothesis.

(c) Part of the problem discussed in (b) is due to the fact that most K-M estimated survival curves are presented without any indication of the uncertainty in the estimate that is due to sampling variability. This imprecision is usually quantified via the standard error – the standard deviation of the sampling distribution associated with the method of estimation. However, a standard error for the estimated survival probability at a particular time t can be calculated, and often appears in a computer listing of the calculations relating to a Kaplan-Meier curve. A range of plausible values for the estimated probability at t units is the estimate plus or minus twice the standard error (see chapter 8). It is essential to indicate an interval such as this one if any important conclusions are to be deduced from the K-M estimate.

A very rough estimate of the standard error is given by Peto et al. [10]. If the estimated survival probability at t units is p and n individuals are still under observation, then the estimated standard error is $p\sqrt{(1-p)/n}$. Since this is an approximate formula, it is possible that the range of plausible values $p \pm 2p\sqrt{(1-p)/n}$ may not lie entirely between 0 and 1; recall that all probabilities fall between these limits. If this overlap represents a serious problem, then it would be wise to consult a statistician.

(d) A critical factor in the calculation of the K-M estimated survival curve is the assumption that the reason an observation has been censored is independent of or unrelated to the cause of death. This assumption is true, for example, if censoring occurs because an individual has only been included in a trial for a specified period of observation and is still being followed. If individuals who responded poorly to a treatment were dropped from a study before death and identified as censored observations, then the K-M estimated survival curve would not be appropriate because the independent censoring assumption has been violated.

There is no good way to adjust for inappropriate censoring so it should, if possible, be avoided. Perhaps the most frequent example of this problem is censoring due to causes of death other than the particular cause which is under study. Unless the different causes of death act independently (and this assumption cannot be tested in most cases), the production of a K-M estimated survival curve corresponding to a specific cause is unwise. Instead, cause-specific estimation techniques that handle the situation appropriately should be used. Although these cause-specific methods are closely related to

Table 6.1. A typical tabulation for a Kaplan-Meier survival curve; the data represent 31 lymphoma patients, of whom 20 are known to have died

Time, t, in months	Number at risk at t months	Number of deaths at t months	Estimated probability of surviving more than t months	Standard error of the estimate
0	31	0	1.000	–
2.5	31	1	0.968	0.032
4.1	30	1	0.935	0.044
4.6	29	1	0.903	0.053
6.4	28	1	0.871	0.060
6.7	27	1	0.839	0.066
7.4	26	1	0.806	0.071
7.6	25	1	0.774	0.075
7.7	24	1	0.742	0.079
7.8	23	1	0.710	0.082
8.8	22	1	0.677	0.084
13.3	21	1	0.645	0.086
13.4	20	1	0.613	0.087
18.3	19	1	0.581	0.089
19.7	18	1	0.548	0.089
21.9	17	1	0.516	0.090
24.7	16	1	0.484	0.090
27.5	15	1	0.452	0.089
29.7	14	1	0.419	0.089
32.9	12	1	0.384	0.088
33.5	11	1	0.349	0.087

those used to produce K-M estimated survivor curves, interpretation of the resulting estimated curves is not straightforward, and statistical help is advised.

(e) Corresponding to the K-M estimated survival curve presented in figure 6.1, table 6.1 shows typical information provided by computer programs which calculate Kaplan-Meier estimated survival curves. Each line of the table records a survival time, the number of deaths which occurred at that time, the number of individuals under observation at that time, the K-M estimate of the probability of surviving beyond that time, and a standard error of the estimate. Typically, the standard error will not be the simple estimate of comment (c), but generally a better one involving more detailed calculations.

6.3. Computing the Kaplan-Meier Estimate

In this penultimate section of chapter 6 we intend to discuss, in some detail, a simple method for calculating the Kaplan-Meier estimate by hand. The technique is not complicated; nevertheless, some readers may prefer to omit this section.

To calculate a Kaplan-Meier estimated survival curve, it is first necessary to list all observations, censored and uncensored, in increasing order. The convention is adopted that if both observed and censored survival times of the same duration have been recorded, then the uncensored observations precede the corresponding censored survival times in the ordered list. For the simple example presented in figure 6.2a, the ordered list of observations, in days, is

1*, 3, 4*, 5, 5, 6*, 7, 7, 7*, 8*

where * indicates a censored survival time.

The second step in the calculation is to draw up a table with six columns labelled as follows:

t observed distinct survival time, in days
n the number of individuals still under observation at t days
r the number of deaths recorded at t days
p_c the proportion of individuals under observation at t days who do not die at t days, i.e., $(n - r)/n$
$\Pr(T > t)$ the estimated probability of survival beyond t days
s.e. the approximate standard error for the estimated probability of survival beyond t days

For convenience, the initial row of the table may be completed by entering $t = 0$, n = number of individuals in the study, $r = 0$, $p_c = 1$, $\Pr(T > t) = 1$ and s.e. = blank. This simply indicates that the estimated survival curve begins with $\Pr(T > 0) = 1$, i.e., the estimated probability of survival beyond Day 0 is one. The first observed survival time in the ordered list of observations is then identified, and the number of observations which exceed this observed survival time is recorded, along with the number of deaths occurring at this specific time. Notice that the number of individuals still under observation at t days is equal to the number of individuals in the study minus the total number of observation times, censored or uncensored, which were less than t days. Thus, both deaths and censored observations reduce the number of individuals still under observation, and therefore at risk of failing, between observed distinct survival times.

In our example, the first observed survival time is $t = 3$; therefore, the initial two rows of the table would now be:

t	n	r	p_c	Pr(T > t)	s.e.
0	10	0	1.0	1.0	–
3	9	1			

The column labelled p_c gives the estimated probability of surviving Day t for individuals alive at t days. This is (n – r)/n, and Pr(T > t), the estimated probability of survival beyond t days, is the product of the value of p_c from the current line and the value of Pr(T > t) from the preceding line of the table. In the second row of our example table then, $p_c = 8/9 = 0.89$ and Pr(T > t) = 0.89 × 1.0 = 0.89. According to the formula given in comment (c) of §6.2 for an approximate standard error for Pr(T > t), the entry in the s.e. column will be:

$$\Pr(T > t)\sqrt{\{1 - \Pr(T > t)\}/n} = 0.89\ \sqrt{(1 - 0.89)/9} = 0.10.$$

This process is then repeated for the next observed survival time. According to the ordered list for our example, r = 2 deaths are recorded at t = 5, at which time n = 7 individuals are still under observation. Therefore, $p_c = (7 - 2)/7 = 0.71$, Pr(T > t) = 0.71 × 0.89 = 0.64, and s.e. = $0.64\sqrt{0.36/7} = 0.15$. The table now looks like the following:

t	n	r	p_c	Pr(T > t)	s.e.
0	10	0	1.0	1.0	-
3	9	1	0.89	0.89	0.10
5	7	2	0.71	0.64	0.15

The final row in the table will correspond to the r = 2 deaths which are recorded at t = 7, when n = 4 individuals are still under observation. In this row, therefore, $p_c = 2/4 = 0.50$, Pr(T > t) = 0.64 × 0.50 = 0.32 and s.e. = $0.32\sqrt{0.68/4} = 0.13$. Thus, the completed table for our simple example is:

t	n	r	p_c	Pr(T > t)	s.e.
0	10	0	1.0	1.0	–
3	9	1	0.89	0.89	0.10
5	7	2	0.71	0.64	0.15
7	4	2	0.50	0.32	0.13

To plot the K-M estimated survival curve from the completed table, use the values in columns one and five (labelled t and Pr(T > t)) to draw the graph with its characteristic staircase appearance. Steps will occur at the values of the distinct observed survival times, i.e., the values of t. To the right of each value of t, draw a horizontal section at a height equal to the corresponding value of Pr(T > t). Each horizontal section will extend from the current value of t to the next value of t, where the next decrease in the graph occurs.

From the columns labelled t and Pr(T > t) in the table for our example, we see that the graph of the K-M estimated survival curve has a horizontal section from t = 0 to t = 3 at probability 1.0, a horizontal section from t = 3 to t = 5 at probability 0.89, a horizontal section from t = 5 to t = 7 at probability 0.64, and a horizontal section from t = 7 at probability 0.32. Since the final entry in the table for Pr(T > t) is not zero, the last horizontal section of the graph is terminated at the largest censored survival time in the ordered list of observations, namely 8. Therefore, the final horizontal section of the graph at probability 0.32 extends from t = 7 to t = 8; the estimated probability of survival is not defined for t > 8 in this particular case.

If the final censored observation, 8, had been an observed survival time instead, then the completed table would have been:

t	n	r	p_c	Pr(T > t)	s.e.
0	10	0	1.0	1.0	–
3	9	1	0.89	0.89	0.10
5	7	2	0.71	0.64	0.15
7	4	2	0.50	0.32	0.13
8	1	1	0.00	0.00	–

In this case, the graph of the K-M estimated survival curve would drop to zero at t = 8 and be equal to zero thereafter.

6.4. A Novel Use of the Kaplan-Meier Estimator

Despite the fact that estimation of the statistical distribution of the time to an endpoint like death is routinely referred to as the survival curve or survivor function, not all studies in which K-M estimators are used require subjects to die or survive. James and Matthews [11] describe a novel use of the K-M estimator in which the endpoint of interest is an event called a donation attempt. Their paradigm, which they call the donation cycle, represents the use of familiar tools like K-M estimator in an entirely new setting, enabling transfusion researchers to study the return behaviour of blood donors quantitatively – something that no one had previously recognized was possible.

According to the paradigm that they introduce, each donation cycle is defined by a sequence of four consecutive events: an initiating donation attempt,

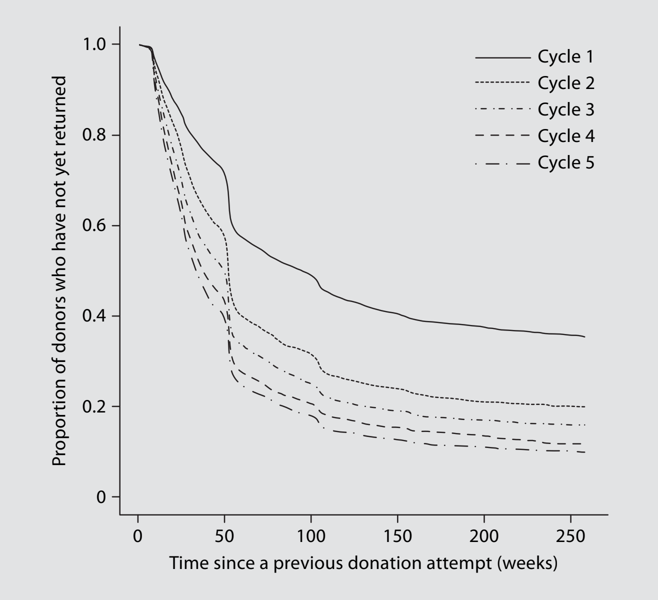


Fig. 6.3. Kaplan-Meier estimated survival curves for Type O, whole blood donations in donation cycles one to five.

a mandatory deferral period, which is typically eight weeks for whole blood donors, an elective interval, and a subsequent donation attempt. The elective interval represents the time between the end of the mandatory deferral period and the occurrence of the next donation attempt. Donors may, and hopefully do, complete multiple donation cycles. Each cycle, while conceptually identical to all other donation cycles, can be distinguished by its unique sequence number, e.g., first, second, etc., and has an exact length which can be measured in suitable units of time, such as days or weeks. Furthermore, if a donor who has made at least one donation attempt fails to make another, the observed interval is clearly a censored observation on the event of interest, i.e., a subsequent donation attempt.

To illustrate the merits of the donation cycle paradigm, James and Matthews [12] obtained records concerning all Type O whole blood donors from the Gulf Coast Regional Blood Center (GCRBC) in Houston, Texas, who had not been permanently deferred before April 15, 1987. From 164,987 usable donor records and a corresponding set of 608,456 transactions that were almost exclusively donation attempts, these researchers identified the lengths of all first, second, third, etc. donation cycles. They then randomly sampled approximately 5,100 intervals from each cycle-specific dataset.

Despite the size of each sample, this cycle-specific data is exactly what modern statistical software requires to generate K-M estimates of the so-called survival curve for each sample. The resulting graphs are displayed in figure 6.3, and illustrate how, for the type O, whole blood donors that contributed through the GCRBC, the intervals between donation attempts depend on a history of previous donations and the elapsed time since the initial or index donation attempt in a cycle. In each case, the estimated probability of making a subsequent donation attempt remains at or near 1.0 for the first eight weeks of the observation period – precisely the length of the mandatory deferral interval – and only thereafter begins to decrease. The graphs also show that the estimated survival function decreases with each additional donation attempt. Thus, for any fixed time exceeding eight weeks, the proportion of donors who had already attempted a subsequent donation is an increasing function of the number of previously completed donation cycles. In contrast to the usual staircase look of most K-M estimated survival curves, the smooth appearance of these estimated functions reflects the effect of using very large sample sizes to estimate, at any fixed value for t, the proportion of donors who had not yet attempted a subsequent donation. One year after the index donation, some 68% of first-time donors had yet to return for a subsequent attempt; at five years, approximately one-third of first-time donors had not attempted a second whole blood donation. The corresponding estimated percentages for donors in their fifth cycle were 36 and 10%, respectively.

The marked localized changes in the graph at 52 and 104 weeks are a particularly distinctive feature of these estimated curves; a third drop appears to occur at 156 weeks, although this decrease is less evident visually. James and Matthews speculate that these parallel changes ‘were the result of two GCRBC activities: blood clinic scheduling and a donor incentive programme which encouraged such clinic scheduling.’ The GCRBC offered an insurance-like scheme that provided benefits for donors and their next-of-kin, provided a donation had been made during the preceding 12 months. Similar localized changes have been seen in studies of other forms of insurance. For example, return-to-work rates usually increase markedly at or near the end of a strike or the expiration of unemployment insurance coverage (see Follmann et al. [13]).

In general, the K-M estimated survival curve is used mainly to provide visual insight into the observed experience concerning the time until an event of interest such as death, or perhaps a subsequent attempt to donate a unit of blood, occurs. It is common, however, to want to compare this observed experience for two or more groups of patients. Although standard errors can be used for this purpose if only a single fixed point in time is of scientific interest, better techniques are available. These include the log-rank test, which we describe in chapter 7.

7

The Log-Rank or Mantel-Haenszel Test for the Comparison of Survival Curves

7.1. Introduction

In medical research, we frequently wish to compare the survival (or relapse, etc.) experience of two groups of individuals. The groups will differ with respect to a certain factor (treatment, age, sex, stage of disease, etc.), and it is the effect of this factor on survival which is of interest. For example, figure 7.1 presents the Kaplan-Meier estimated survival curves for the 31 lymphoma patients mentioned in chapter 6, and a second group of 33 lymphoma patients who were diagnosed without clinical symptoms. According to standard practice, the 31 patients with clinical symptoms are said to have 'B' symptoms, while the other 33 have 'A' symptoms. Figure 7.1 shows an apparent survival advantage for patients with A symptoms.

In the discussion that follows, we present a test of the null hypothesis that the survival functions for patients with A and B symptoms are the same, even though their respective Kaplan-Meier estimates, which are based on random samples from the two populations, will inevitably differ. This test, which is called the log-rank or Mantel-Haenszel test, is only one of many tests which could be devised; nevertheless, it is frequently used to compare the survival functions of two or more populations.

The log-rank test is designed particularly to detect a difference between survival curves which results when the mortality rate in one group is consistently higher than the corresponding rate in a second group and the ratio of these two rates is constant over time. This is equivalent to saying that, provided an individual has survived for t units, the chance of dying in a brief interval following t is k times greater in one group than in the other, and the same statement is true for all values of t. The null hypothesis that there is no differ-

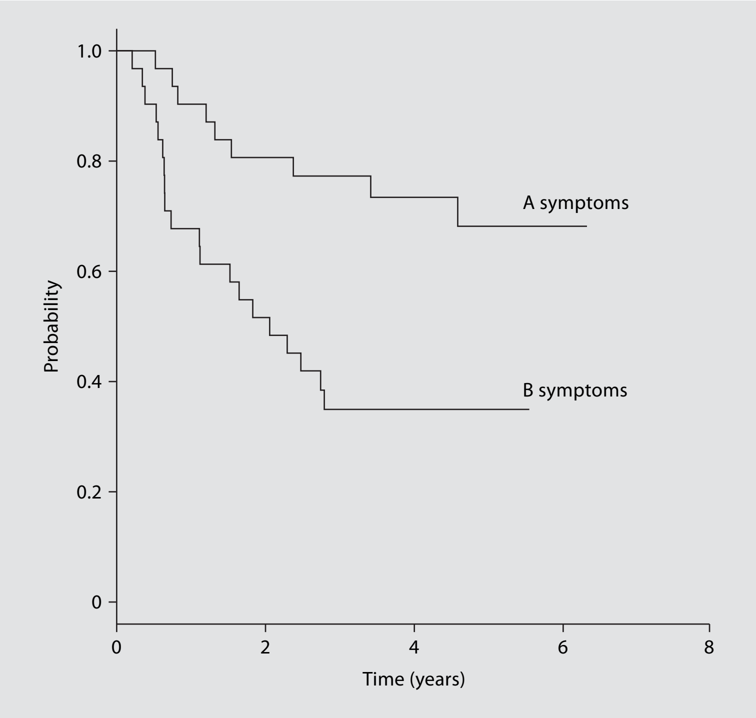


Fig. 7.1. The Kaplan-Meier estimated survival curves for 33 lymphoma patients presenting with A symptoms and 31 with B symptoms.

ence in survival experience between the two groups is represented by the value k = 1, i.e., a ratio of one. As we indicated in chapter 6, time is usually measured from some well-defined event such as diagnosis.

7.2. Details of the Test

The basic idea underlying the log-rank test involves examining each occasion when one or more deaths (or events) occurs. Based on the number of individuals in each group who are alive just before the observed death time and the total number of deaths observed at that time, we can calculate how many deaths would be expected in each group if the null hypothesis is true, i.e., if the mortality rates are identical. For example, if Group 1 has six individuals alive at t units and Group 2 has three, then the observed deaths at t should be distributed in the ratio 2:1 between the two groups, if the null hypothesis is true.

If three deaths actually occurred at t units, then we would expect two in the first group and one in the second group. If only one death had actually occurred at t, then we would say that the expected number of deaths in Group 1 is 2/3 and in Group 2 is 1/3. Notice that the expected number of deaths need not correspond to a positive integer.

To complete the log-rank test we add up, for the two groups separately, the observed and expected numbers of deaths at all observed death times. These numbers are then compared. If O_1 and O_2 are the observed numbers of deaths in the two groups and E_1 and E_2 are the expected numbers of deaths calculated by summing the expected numbers at each event time, then the statistic used for comparison purposes is

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}.$$

If the null hypothesis is true, T should be distributed approximately as a χ_1^2 random variable (chi-squared with one degree of freedom). Let t_o represent the observed value of T for a particular set of data; then the significance level of the log-rank test is given by $\Pr(T \geq t_o)$, which is approximately equal to $\Pr(\chi_1^2 \geq t_o)$. Therefore, we can use table 4.10 to evaluate the significance level of the log-rank test (see §4.4).

Comments:

(a) If we wish to compare the survival experience in two groups specified, say, by treatment, but it is important to adjust the comparison for another prognostic factor, say stage of the disease, then a stratified log-rank test may be performed. In this case, study subjects must be classified into strata according to stage, and within each stratum the calculations of observed and expected numbers of deaths for a log-rank test are performed. The test statistic, T, is computed from values of O_1, O_2, E_1 and E_2 which are obtained by summing the corresponding observed and expected values from all the strata.

The effectiveness of stratification as a means of adjusting for other prognostic factors is limited because it is necessary, simultaneously, to retain a moderate number of subjects in each stratum. If we wish to adjust for a number of prognostic factors then, in principle, we can define prognostic strata within which prognosis would be similar, except for the effect of treatment. However, as the number of prognostic factors increases, the strata soon involve too few subjects to be meaningful. Unless the study is very large, the use of more than six or eight strata is generally unwise. Stratification is probably most effective with two to four strata, especially since there are procedures which are more useful when the number of prognostic factors is large. These procedures will be discussed in chapter 13.

(b) It is possible to restrict the comparison of survival experience in two groups to a specified interval of the observation period, since the log-rank test is still valid when it is restricted in this way. However, it is important not to choose such restricted time intervals by examining the observed data and selecting an interval where the mortality looks different; this method of choosing an interval invalidates the calculation of the p-value because it constitutes selective sampling from the observed survival experience. On the other hand, it would be very reasonable to examine, separately, early and late mortality in the two treatment groups, if the time periods early and late are defined in some natural way.

(c) The log-rank test is always a valid test of the null hypothesis that the survival functions of two populations are the same. The effectiveness of the test in detecting departures from this hypothesis does depend on the form of the difference. The log-rank test is, in some sense, optimal if the difference arises because the mortality rate in one group is a constant multiple of the corresponding rate in the other group. If this is not the case, then the log-rank test may not be able to detect a difference that does exist. The importance of this assumption that the ratio of mortality rates is constant with respect to time is another reason for performing the log-rank test on suitably-defined, restricted time intervals; doing so helps to validate the constant ratio assumption for the data at hand.

(d) There are a number of alternative tests for comparing the survival experience of two groups. We do not intend to discuss these tests except to mention, briefly, another commonly-used test called the generalized Wilcoxon test. The latter differs from the log-rank test in that it attaches more importance to early deaths than to later deaths, whereas the log-rank test gives equal weight to all deaths. Thus, differences in the early survival experience of two populations are more likely to be detected by the generalized Wilcoxon test than by the log-rank test which we have described.

If survival data include censored observations, then there are different versions of the generalized Wilcoxon test which may be used. Although the details of their research are beyond the scope of this book, the work of Prentice and Marek [14] suggests that Gehan's [15] generalization is subject to serious criticism and probably should be avoided. An alternative, called Peto's generalized Wilcoxon statistic (see Peto and Peto [16]), is preferable in this case.

(e) The use of $T = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$ as the test statistic, and the assumption that T has a χ_1^2 distribution if the null hypothesis is true, is an approximation which is particularly convenient for hand calculations. This approximation can be improved and, frequently, computer programs to calculate the log-rank test will use an alternative form of the statistic. We will not discuss any of these improved versions, except to note that the principle and nature of the test are unchanged.

Table 7.1. Hypothetical survival times, in days, for the comparison of mortality in two groups; a * indicates a censored observation

Group 1	1*, 3, 4*, 5, 5, 6*, 7, 7, 7*, 8
Group 2	2, 2, 3*, 4, 6*, 6*, 7, 10

(f) The log-rank test can be generalized to test for the equality of survival experience in more than two study populations (groups). The total number of deaths at each event time, and the proportions of the study subjects in each group at that time, are used to calculate the expected numbers of deaths in each group if the null hypothesis is true. The totals O and E are calculated for each group, and the test statistic, T, is the sum of the quantity $(O - E)^2/E$ for each group. If the null hypothesis is true, T is distributed approximately as a χ^2_{k-1} variable (chi-squared with k – 1 degrees of freedom), where k is the number of groups whose survival experience is being compared.

7.3. Evaluating the Log-Rank Test – A Simple Example

Although statisticians will frequently use statistical software to evaluate the log-rank test for a given set of data, the actual calculations are not particularly complicated. To illustrate exactly what is involved, we consider the simple example shown in table 7.1. This table presents two sets of survival times; those in Group 1 correspond to the times used in the example discussed in §6.3. To perform the calculations required for a log-rank test, it is convenient to set up a table with ten columns. The column headings will be the following:

t	the event time, in days
n	the number of individuals still under observation at t days
n_1	the number of individuals in Group 1 still under observation at t days
n_2	the number of individuals in Group 2 still under observation at t days
r	the number of deaths recorded at t days
c	the number of censored values recorded at t days
o_1	the number of deaths in Group 1 recorded at t days
o_2	the number of deaths in Group 2 recorded at t days
e_1	the expected number of deaths in Group 1 at t days
e_2	the expected number of deaths in Group 2 at t days

Chronologically, the first event recorded in table 7.1 is a censored value occurring at t = 1. Although this event actually contributes nothing to our test statistic, it is convenient, for completeness, to include a row in the table for this censored value. The event time is t = 1, the total number of observations is n =

18, of which $n_1 = 10$ and $n_2 = 8$ are in Groups 1 and 2, respectively. At $t = 1$ there are $r = 0$ deaths and $c = 1$ censored values. Since no deaths were observed, the observed and expected numbers of deaths are all zero as well. Therefore, the first row of the table is:

t	n	n_1	n_2	r	c	o_1	o_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0

The second event time in our hypothetical study is $t = 2$, at which two deaths were recorded in Group 2. One observation was censored at $t = 1$; therefore, at $t = 2$ we have $n = 17$ individuals under observation, of which $n_1 = 9$ and $n_2 = 8$ are in Groups 1 and 2, respectively. The number of deaths is $r = 2$ and the number of censored observations is $c = 0$. The observed numbers of deaths are $o_1 = 0$ in Group 1 and $o_2 = 2$ in Group 2. The proportion of individuals under observation in Group 1 is $\frac{9}{17}$ and in Group 2 it is $\frac{8}{17}$. Therefore, the expected numbers of deaths in the two groups, given that there were two observed deaths, are $e_1 = 2 \times \frac{9}{17} = 1.06$ and $e_2 = 2 \times \frac{8}{17} = 0.94$. Notice that, except for rounding errors in the calculations, it should be true that $r = o_1 + o_2 = e_1 + e_2$. With the second row completed the table now becomes

t	n	n_1	n_2	r	c	o_1	o_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0
2	17	9	8	2	0	0	2	1.06	0.94

At the third event time, $t = 3$, there is one death recorded in Group 1 and one censored observation in Group 2. At $t = 3$ there were $n = 15$ individuals still under observation, of which $n_1 = 9$ and $n_2 = 6$ were in Groups 1 and 2, respectively. Notice that, in each row of the table, the current value of n is the value of n from the preceding row minus the sum of r and c from the preceding row. At $t = 3$ there was $r = 1$ death and $c = 1$ censored value with $o_1 = 1$ and $o_2 = 0$; therefore, the expected numbers of deaths are $e_1 = 1 \times \frac{9}{15} = 0.60$ and $e_2 = 1 \times \frac{6}{15} = 0.40$.

Additional rows are added to the table until the last observed death is recorded, or until n_1 or n_2 becomes zero. The completed table for the example is given in figure 7.2.

When the table has been completed, the last four columns must be summed to obtain $O_1 = \Sigma\, o_1 = 6$, $O_2 = \Sigma\, o_2 = 4$, $E_1 = \Sigma\, e_1 = 6.05$ and $E_2 = \Sigma\, e_2 = 3.95$. Since the observed value of T, which is calculated in figure 7.2, is $t_0 = 0.001$, the significance level of the data with respect to the null hypothesis is given by

Column Headings

t	the event time, in days
n	the number of individuals still under observation at t days
n_1	the number of individuals in Group 1 still under observation at t days
n_2	the number of individuals in Group 2 still under observation at t days
r	the number of deaths recorded at t days
c	the number of censored values recorded at t days
o_1	the number of deaths in Group 1 recorded at t days
o_2	the number of deaths in Group 2 recorded at t days
e_1	the expected number of deaths in Group 1 at t days
e_2	the expected number of deaths in Group 2 at t days

t	n	n_1	n_2	r	c	o_1	o_2	$e_1 = \frac{r \times n_1}{n}$	$e_2 = \frac{r \times n_2}{n}$
1	18	10	8	0	1	0	0	$\frac{0 \times 10}{18} = 0.00$	$\frac{0 \times 8}{18} = 0.00$
2	17	9	8	2	0	0	2	$\frac{2 \times 9}{17} = 1.06$	$\frac{2 \times 8}{17} = 0.94$
3	15	9	6	1	1	1	0	$\frac{1 \times 9}{15} = 0.60$	$\frac{1 \times 6}{15} = 0.40$
4	13	8	5	1	1	0	1	$\frac{1 \times 8}{13} = 0.62$	$\frac{1 \times 5}{13} = 0.38$
5	11	7	4	2	0	2	0	$\frac{2 \times 7}{11} = 1.27$	$\frac{2 \times 4}{11} = 0.73$
6	9	5	4	0	3	0	0	$\frac{0 \times 5}{9} = 0.00$	$\frac{0 \times 4}{9} = 0.00$
7	6	4	2	3	1	2	1	$\frac{3 \times 4}{6} = 2.00$	$\frac{3 \times 2}{6} = 1.00$
8	2	1	1	1	0	1	0	$\frac{1 \times 1}{2} = 0.50$	$\frac{1 \times 1}{2} = 0.50$
Totals						6	4	6.05	3.95

$$t_o = \frac{(6-6.05)^2}{6.05} + \frac{(4-3.95)^2}{3.95} = 0.001$$

Fig. 7.2. Details of the log-rank test calculations for the data presented in table 7.1.

Table 7.2. Survival times, in months, for 64 lymphoma patients; a * indicates a censored survival time

A symptoms	3.2*, 4.4*, 6.2, 9.0, 9.9, 14.4, 15.8, 18.5, 27.6*, 28.5, 30.1*, 31.5*, 32.2*, 41.0, 41.8*, 44.5*, 47.8*, 50.6*, 54.3*, 55.0, 60.0*, 60.4*, 63.6*, 63.7*, 63.8*, 66.1*, 68.0*, 68.7*, 68.8*, 70.9*, 71.5*, 75.3*, 75.7*
B symptoms	2.5, 4.1, 4.6, 6.4, 6.7, 7.4, 7.6, 7.7, 7.8, 8.8, 13.3, 13.4, 18.3, 19.7, 21.9, 24.7, 27.5, 29.7, 30.1*, 32.9, 33.5, 35.4*, 37.7*, 40.9*, 42.6*, 45.4*, 48.5*, 48.9*, 60.4*, 64.4*, 66.4*

$\Pr(\chi_1^2 \geq 0.001)$. According to table 4.10, $\Pr(\chi_1^2 \geq 0.001) > 0.25$. Therefore, we conclude that the data provide no evidence to contradict the null hypothesis that the survival functions for Groups 1 and 2 are the same.

7.4. More Realistic Examples

Table 7.2 presents the data for the two Kaplan-Meier estimated survival curves shown in figure 7.1. The observed numbers of deaths in the A and B symptoms groups were 9 and 20, respectively. The log-rank calculations lead to corresponding expected numbers of deaths of 17.07 and 11.93. Therefore, the observed value of the log-rank test statistic for these data is

$$t_0 = \frac{(9-17.07)^2}{17.07} + \frac{(20-11.93)^2}{11.93} = 9.275.$$

If the null hypothesis of no difference in survival experience between the two groups is true, t_o should be an observation from a χ_1^2 distribution. Therefore, the significance level of the data is equal to

$$\Pr(\chi_1^2 \geq 9.275) = 0.0023.$$

Using table 4.10, we can only determine that $0.005 > \Pr(\chi_1^2 \geq 9.275) > 0.001$; the exact value was obtained from other sources. In either case, we would conclude that the data do provide evidence against the null hypothesis and suggest that lymphoma patients with A symptoms enjoy a real survival advantage.

Notice that the essence of the test is to conclude that we have evidence contradicting the null hypothesis when the observed numbers of deaths in the two groups are significantly different from the corresponding numbers which would be expected, if the survival functions for the two groups of patients are the same.

The data on the return behaviour of type O whole blood donors during the first five donation cycles that are summarized visually in the estimated curves displayed in figure 6.3 are too numerous to tabulate here, since each of the five estimates is based on a random sample of approximately 5,100 donors. According to James and Matthews [12], the observed numbers of cycles completed within 260 weeks of an index donation attempt were 2,812, 3,551, 3,678, 3,888 and 3,975 for individuals belonging to donation cycles 1 through 5, respectively. In view of the very large sample sizes involved, there is no doubt, statistically, that these five estimated donor return functions differ markedly from each other. In this instance, a log-rank test to investigate the evidence in the data concerning the usual hypothesis that the five donor return functions are the same is unnecessary. Because of the sample sizes and the distinct differences among the estimated curves, the observed value of the log-rank test statistic is enormous, and the corresponding significance level of the data – which would be evaluated using χ^2_4 distribution – is 0. The data represent unequivocal evidence that a history of prior donation attempts is associated with substantially shorter intervals between subsequent attempts to donate type O whole blood.

In subsequent chapters, we intend to present important extensions and generalizations of the methods we have already discussed. However, we first need to introduce certain notions and concepts which are based on the normal, or Gaussian, distribution. In general, statistical methods for normal data comprise a major fraction of the material presented in elementary textbooks on statistics. Although the role of this methodology is not as prominent in clinical research, the techniques still represent an important set of basic, investigative tools for both the statistician and the medical researcher. In addition, the normal distribution plays a rather important part in much of the advanced methodology that we intend to discuss in chapters 10 to 15. Therefore, in chapter 8, we will introduce the normal distribution from that perspective. Chapter 9 contains details of the specific methods which can be used with data which are assumed to be normally distributed.

8

An Introduction to the Normal Distribution

8.1. Introduction

In chapter 1, the normal distribution was used to describe the variation in systolic blood pressure. In most introductory books on statistics, and even in many advanced texts, the normal probability distribution is prominent. By comparison, in the first seven chapters of this book the normal distribution has been mentioned only once. This shift in emphasis has occurred because the specific test procedures which are appropriate for data from a normal distribution are not used as frequently in medical research as they are in other settings such as the physical sciences or engineering. In fact, the nature of medical data often precludes the use of much of the methodology which assumes that the data are normally distributed. Nevertheless, there are certain situations where these techniques will be useful, and in chapter 9 we discuss a number of the specialized procedures which can be used to analyze normally distributed data.

There is another way that the normal distribution arises in medical statistics however, and for this reason, alone, it is essential to understand the basic theory of the normal distribution. Many of the advanced statistical methods that are being used in medical research today involve test statistics which have approximate normal distributions. This statement applies particularly to the methods which we intend to discuss in chapters 10 through 15. Therefore, in the remainder of this chapter, we shall introduce the normal distribution. We have not previously discussed any probability distribution in quite as much detail as our scrutiny of the normal distribution will involve. Therefore, readers may find it helpful to know that a complete appreciation of §8.2 is not essential. The most important material concerns aspects of the normal distribution which are pertinent to its use in advanced methodology, and these topics are discussed in §§8.3, 8.4. However, §8.2 should be read before proceeding to the rest of the chapter.

8.2. Basic Features of the Normal Distribution

The normal distribution is used to describe a particular pattern of variation for a continuous random variable. Our discussion of continuous random variables in §1.2 emphasized the area = probability equation, and introduced the cumulative probability curve as a means of calculating $\Pr(X \le a)$, the probability that the random variable X is at most a. Equivalently, $\Pr(X \le a)$ is the area under the probability curve for X which lies to the left of the vertical line at a.

Figure 8.1a shows the cumulative probability curve for a continuous random variable, Z, which has a *standardized* normal distribution. The upper case Roman letter Z is generally used to represent the standardized normal variable, and we will follow this convention in the remainder of the book.

The cumulative probability curve is very useful for calculating probabilities; however, the graph in figure 8.1a does not convey particularly well the nature of the variation described by the standardized normal distribution. To appreciate this aspect of Z we need to examine its probability curve, shown in figure 8.1b. Recall that the probability curve is the analogue of a histogram for a continuous random variable. Immediately we see that the standard normal distribution has a characteristic bell-like shape and is symmetric about the mean value, zero. Because of this symmetry about zero, it follows that for any positive number a, the probability that Z exceeds a is equal to the probability that Z is less than –a, i.e., $\Pr(Z > a) = \Pr(Z < -a)$. This equality is indicated in figure 8.1b by the two shaded areas of equal size which represent these two probabilities.

In chapter 1, we described the mean, $E(X) = \mu$, and the variance, $\text{Var}(X) = \sigma^2$, as useful summary characteristics of a probability distribution. For the standardized normal distribution the mean is zero and the variance is one; that is, $\mu = 0$ and $\sigma^2 = 1$. These are, in fact, the particular characteristics which define the standardized normal distribution. Now, compare figures 8.1a and 8.2a. Figure 8.2a shows the cumulative probability curve for a random variable, X, which is also normally distributed but which has mean μ and variance σ^2, i.e., $E(X) = \mu$, $\text{Var}(X) = \sigma^2$. The curve for X corresponds exactly with the cumulative probability curve for Z, shown in figure 8.1a, except that the center of the horizontal axis in figure 8.2a is μ rather than zero, and each major division on this axis represents one standard deviation, σ, rather than one unit. Clearly, if $\mu = 0$ and $\sigma = 1$, then X and Z would have identical cumulative probability curves. This is why the two values $E(Z) = 0$ and $\text{Var}(Z) = 1$, i.e., $\mu = 0, \sigma^2 = 1$, characterize the standardized normal distribution among all possible normal distributions.

A convenient shorthand for specifying that X has a normal distribution with mean μ and variance σ^2 is the notation $X \sim N(\mu, \sigma^2)$, which we will use

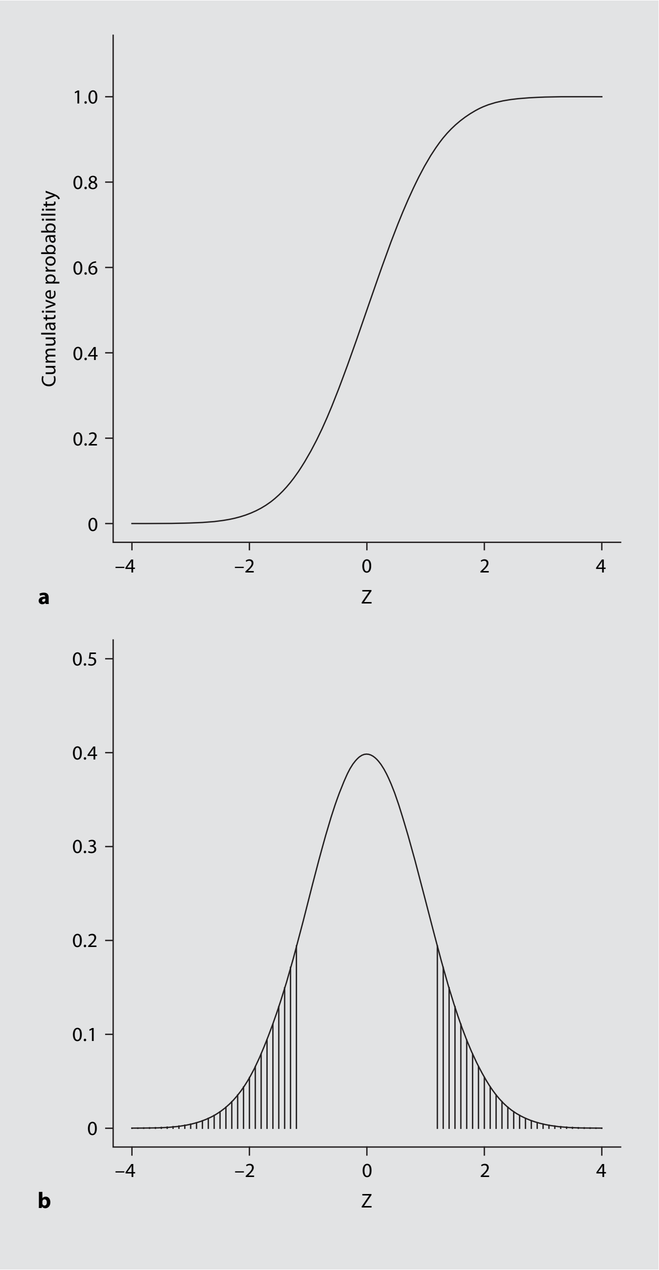


Fig. 8.1. The standardized normal distribution. **a** Cumulative probability curve. **b** Corresponding probability curve with shaded areas representing Pr(Z > a) and Pr(Z < –a).

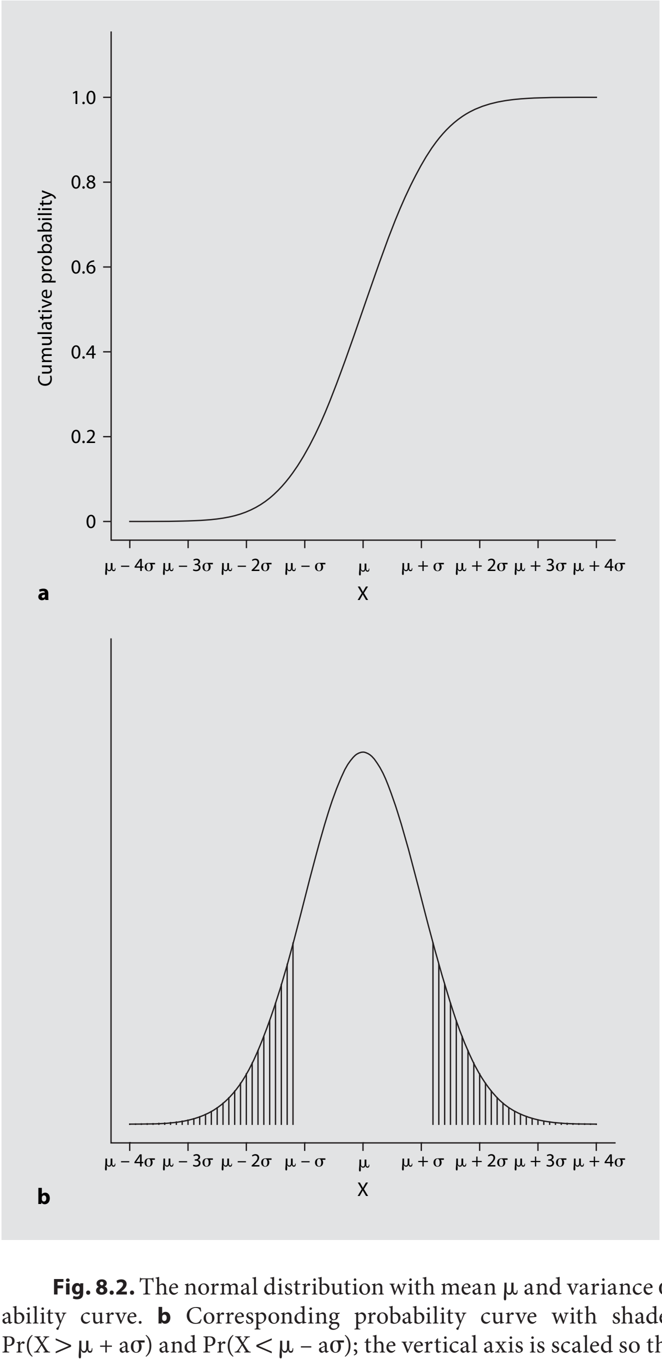


Fig. 8.2. The normal distribution with mean μ and variance σ^2. **a** Cumulative probability curve. **b** Corresponding probability curve with shaded areas representing $\Pr(X > \mu + a\sigma)$ and $\Pr(X < \mu - a\sigma)$; the vertical axis is scaled so that the total area under the curve is one.

repeatedly in chapters 8 and 9. To indicate that Z is a standardized normal random variable we simply write $Z \sim N(0, 1)$.

An important fact which we will often use is the following:

$$\text{if } X \sim N(\mu, \sigma^2), \quad \text{then } \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Therefore, the probability distribution of $(X - \mu)/\sigma$ is exactly the same as the probability distribution of Z, and so we write $Z = (X - \mu)/\sigma$. The formula asserts that Z is the separation, measured in units of σ, between X and μ; the sign of Z indicates whether X lies to the left ($Z < 0$) or right ($Z > 0$) of μ. This relationship is best appreciated by comparing the probability curves for $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$. Notice that the shape of the curve for X, shown in figure 8.2b, corresponds exactly with the curve for Z shown in figure 8.1b, except that the center of symmetry is located at μ rather than at zero, and each major division on the horizontal axis is one standard deviation, σ, rather than one unit. In figure 8.2b, the equal probabilities which are a result of the symmetry about μ (see the shaded areas) correspond to the probabilities $\Pr(X > \mu + a\sigma)$ on the right and $\Pr(X < \mu - a\sigma)$ on the left. And by comparing figures 8.1b and 8.2b we can also see that

$$\Pr(Z > a) = \Pr(X > \mu + a\sigma) \text{ and } \Pr(Z < -a) = \Pr(X < \mu - a\sigma).$$

This equality between probabilities for $X \sim N(\mu, \sigma^2)$ and probabilities for $Z \sim N(0, 1)$ means that probabilities for X can be calculated by evaluating the equivalent probabilities for the standardized normal variable Z. In fact, since $\Pr(Z > a) = \Pr(X > \mu + a\sigma)$, it follows that

$$\begin{aligned} \Pr(Z > a) &= \Pr(X > \mu + a\sigma) \\ &= \Pr(X - \mu > a\sigma) && \text{[subtract } \mu \text{ from both sides of the inequality]} \\ &= \Pr\left(\frac{X - \mu}{\sigma} > a\right). && \text{[divide both sides of the inequality by } \sigma\text{]} \end{aligned}$$

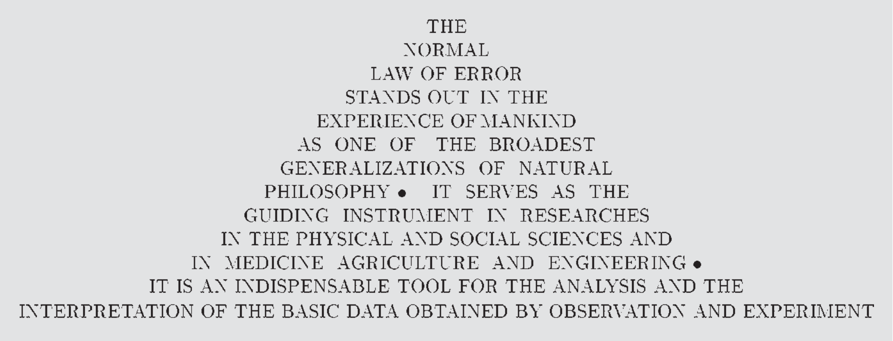
This illustrates, once more, the fact that

$$\frac{X - \mu}{\sigma} = Z \sim N(0, \ 1),$$

i.e., the random variable $\frac{X-\mu}{\sigma}$ has a standardized normal distribution. This relationship, $Z = \frac{X-\mu}{\sigma}$, is usually called the standardizing transformation because it links the distribution of $X \sim N(\mu, \sigma^2)$ to the distribution of $Z \sim N(0, 1)$. In fact, the concept of the standardizing transformation is fundamental to the calculation of normal probabilities and to the use of the normal distribution in the advanced methodology we intend to discuss in chapters 10 through 15.

In §8.1, we indicated that the nature of medical data often precludes the use of methods which assume that the observations, e.g., survival times, are normally distributed. Why, then, is the normal distribution so important? A complete answer to this question goes well beyond the scope of our discussion. However, in general terms we can state that, according to a certain theorem in mathematical statistics called the central limit theorem, the probability distribution of the sum of observations from any population corresponds more and more to that of a normal distribution as the number of observations in the sum increases, i.e., if the sample size is large enough, the sum of observations from any distribution is approximately normally distributed. Since many of the test statistics and estimating functions which are used in advanced statistical methods can be represented as just such a sum, it follows that their approximate normal distributions can be used to calculate probabilities when nothing more exact is possible.

To counteract the solemn tone of the preceding explanation, and also to summarize our discussion of the basic features of the normal distribution, we conclude with a more light-hearted portrait of the normal probability curve which was devised by W.J. Youden:



THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY • IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING •
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

8.3. The Normal Distribution and Significance Testing

In this section and the next, we will focus attention on the role of the normal distribution in the more advanced statistical methods which we intend to introduce in chapters 10 through 15.

Recall the survival data for 64 patients with lymphoma which were discussed in §7.4. In chapter 7, we used a log-rank test to study the possible survival advantage that patients presenting with A symptoms might enjoy relative to those who present with B symptoms. Since the significance level of the log-rank test was 0.0023, we concluded that the data provide evidence against the hypothesis of comparable survival in the two groups.

In chapter 13, we discuss a method for estimating the ratio of the separate mortality rates for patients presenting with A and B symptoms. This method leads to an estimate that the mortality rate for patients with B symptoms is 3.0 times that of patients with A symptoms. The actual calculations which are involved concern the logarithm of this ratio of mortality rates, which we represent by b, and lead to an estimate of b which we denote by $\hat{b} = \log 3.0 = 1.10$. Simultaneously, the calculations also yield $\hat{\sigma}$, the estimated standard deviation of $\hat{b}$; the value of $\hat{\sigma}$ is 0.41. The symbol ˆ indicates that $\hat{b}$ is an estimate of b and $\hat{\sigma}$ is an estimate of the standard deviation of $\hat{b}$. We will also use the phrase est. standard error ($\hat{b}$) interchangeably with $\hat{\sigma}$.

Much of the methodology in chapters 10–15 is concerned with quantities like b. The usual question of interest is whether or not b equals zero, since this value frequently represents a lack of association between two variables of interest in a study. For example, in the lymphoma data, the value b = 0 corresponds to a ratio of mortality rates which is equal to $e^0 = 1.0$. Thus, b = 0 represents the hypothesis that there is no relationship between presenting symptoms and survival. In this section, we will consider a test of the hypothesis H: b = 0 which is based on the quantities $\hat{b}$ and $\hat{\sigma}$.

Because of the central limit theorem, which we discussed briefly at the end of §8.2, we can state that $\hat{b}$, which is usually a complicated function of the data, has an approximate normal distribution with mean b and variance $\hat{\sigma}^2$, i.e., $\hat{b} \sim N(b, \hat{\sigma}^2)$. We can remark here, without any attempt at justification, that one reason for using the logarithm of the ratio of mortality rates, rather than the ratio itself, is that the logarithmic transformation generally improves the accuracy of the normal approximation. Since $\hat{b} \sim N(b, \hat{\sigma}^2)$, to test H: b = 0 we assume that $\hat{b} \sim N(0, \hat{\sigma}^2)$, i.e., assume that the hypothesis is true, and evaluate the degree to which an observed value of $\hat{b}$, say b_o, is consistent with this particular normal distribution. If $\hat{b} \sim N(0, \hat{\sigma}^2)$, then the standardizing transformation ensures that

$$Z = \frac{\hat{b} - 0}{\hat{\sigma}} = \frac{\hat{b}}{\hat{\sigma}} \sim N(0, 1).$$

In general, both positive and negative values of $\hat{b}$ can constitute evidence against the hypothesis that b is zero. This prompts us to use the test statistic

$$T = \frac{|\hat{b}|}{\hat{\sigma}} = \frac{|\hat{b}|}{\text{est. standard error}\,(\hat{b})},$$

to test H: b = 0; recall that $|\hat{b}|$ means that we should change the sign of $\hat{b}$ if its value is negative. Let t_o be the observed value of T which corresponds to b_o, i.e., $t_o = |b_o|/\hat{\sigma}$. Since T is always non-negative, values which exceed t_o are less consistent with the null hypothesis than the observed data. Therefore, the signifi-

Table 8.1. Critical values of the probability distribution of $T = |Z|$; the table specifies values of the number t_o such that $\Pr(T \geq t_o) = p$

	Probability level, p				
	0.10	0.05	0.01	0.005	0.001
t_o	1.645	1.960	2.576	2.807	3.291

Adapted from: Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol l, pp 117–118. It appears here with the kind permission of the publishers.

cance level of the test is $\Pr(T \geq t_o)$, which is equal to $\Pr(|Z| \geq t_o)$ since $T = |Z|$.

Although $\Pr(T \geq t_o)$ can be calculated exactly, it is more common in the medical literature to compare t_o with selected critical values for the distribution of T. These can be derived from the cumulative probability curve for Z, and are tabulated in table 8.1. This use of critical values for the probability distribution of T in order to determine the significance level of the data corresponds exactly with our use of statistical tables for the χ^2 distribution in chapters 4, 5 and 7.

Although the distinction between the random variable $\hat{b}$ and its observed value b_o is technically correct, this difference is often not maintained in practice. As we have indicated earlier, the results in the case of the lymphoma data would be summarized as $\hat{b} = 1.10$ and $\hat{\sigma} = 0.41$. Therefore, the observed value of the test statistic, which is often called a normal deviate, is

$$\frac{\hat{b}}{\hat{\sigma}} = \frac{1.10}{0.41} = 2.68,$$

and by referring to table 8.1 we can see that the significance level of the lymphoma data with respect to the hypothesis $b = 0$ is between 0.01 and 0.005, i.e., $0.005 < p < 0.01$. This indicates that the data provide strong evidence against the null hypothesis that b equals zero, i.e., against the hypothesis of identical survival experience in the two patient groups. We reached the same conclusion in §7.4 after using a log-rank test to analyze the survival times.

In §8.5, we describe how tables of the standardized normal distribution can be used to calculate p-values more precisely. At this point, however, it seems appropriate to indicate the link between the χ^2_1 distribution, which was prominent in previous chapters, and the probability distributions of Z and $T = |Z|$. It can be shown that $T^2 = Z^2 \sim \chi^2_1$; therefore,

$$\Pr(T \geq t_o) = \Pr(T^2 \geq t_o^2) = \Pr(Z^2 \geq t_o^2) = \Pr(\chi^2_1 \geq t_o^2).$$

This equation specifies that statistical tables for the χ_1^2 distribution can also be used to evaluate the significance level of an observed Z-statistic, i.e., an observed value of T. To use the χ_1^2 table, we first need to square t_o. Occasionally, this relationship between T and the χ_1^2 distribution is used to present the results of an analysis. In general, however, significance levels are calculated from observed Z-statistics by referring to critical values for the distribution of T = $|Z|$, cf., table 8.1.

8.4. The Normal Distribution and Confidence Intervals

In all the statistical methods which we have thus far considered, the use of significance tests as a means of interpreting data has been emphasized. The concept of statistical estimation was introduced in chapter 6, but has otherwise been absent from our discussion. Although significance tests predominate in the medical literature, we can discern a shift towards the greater use of estimation procedures, the most useful of which incorporate the calculation of confidence intervals. From the statistical point of view, this is a development which is long overdue.

The perspective which we intend to adopt in this section is a narrow one; a definitive exposition of the topic of confidence intervals goes beyond the purpose of this book. There are well-defined methods for calculating confidence intervals in all of the situations that we have previously described; we shall not consider any of these techniques, although the calculations are generally not difficult. In chapter 9, we will introduce the calculation of a confidence interval for the mean of a population when the data are normally distributed. This particular interval arises naturally in our discussion of specialized procedures for data of this kind. However, in this section we will consider only those confidence intervals which pertain to the role of the normal distribution in the advanced statistical methods which are the subject of chapters 10 through 15. In this way, our discussion will parallel that of the preceding section.

In chapter 2, we defined a significance test to be a statistical procedure which determines the degree to which observed data are consistent with a specific hypothesis about a population. Frequently, the hypothesis concerns the value of a specific parameter which, in some sense, characterizes the population, e.g., μ or σ^2. By suitably revising the wording of this definition, we can accurately describe one method of obtaining confidence intervals. Basically, a confidence interval can be regarded as the result of a statistical procedure which identifies the set of all plausible values of a specific parameter. These will be values of the parameter which are consistent with the observed data. In fact, a confidence interval consists of every possible value of the parameter which, if tested as a specific null hypothesis in the usual way, would not lead us to conclude that the data contradict that particular null hypothesis.

Suppose, as in the previous section, that $\hat{b}$ is an estimate of a parameter b, generated by one of the methods described in chapters 10 through 15; we have stated that the distribution of $\hat{b}$ is approximately normal with mean b and variance $\hat{\sigma}^2$, i.e., $\hat{b} \sim N(b, \hat{\sigma}^2)$. Now imagine that we wish to test the hypothesis that b is equal to the specific value b_1, say. The appropriate test statistic is $T = |\hat{b} - b_1|/\hat{\sigma}$, and if the observed value of T which we obtain is less than 1.96 (cf., table 8.1), then the corresponding significance level exceeds 0.05 since $\Pr(T \geq 1.96) = 0.05$. Consequently, we would conclude that the value b_1 is plausible, since the hypothesis H: $b = b_1$ is not contradicted by the data ($p > 0.05$). Now if b_1 is the true value of b, we know that $\Pr(T \leq 1.96) = 0.95$. But we can also interpret this probability statement in the following way: the probability that our interval of plausible values will include the specific value b_1 is 0.95, if b_1 is, in fact, the true value of b. For this reason, the interval of plausible values is called a 95% confidence interval; we are 95% sure, or confident, that the true value of b will be included in the interval of values which are consistent with the data.

To actually carry out a significance test for each possible value of b would be an enormous task. Fortunately, the actual calculations only need to be performed once, using algebraic symbols rather than actual numbers. This calculation determines that the set of all possible values of the parameter which are consistent with the data, i.e., $p > 0.05$, is the interval $(\hat{b} - 1.96\hat{\sigma}, \hat{b} + 1.96\hat{\sigma})$. Notice that this interval is completely determined by the data, which are summarized in the values of $\hat{b}$ and $\hat{\sigma}$, and by the value 1.96 from the normal distribution, corresponding to the 5% critical value which we use to judge consistency. Since we are 95% sure that this interval includes the true value of b, $(\hat{b} - 1.96\hat{\sigma}, \hat{b} + 1.96\hat{\sigma})$ is called a 95% confidence interval for b.

Although 95% confidence intervals are the most common, since they correspond to the familiar 5% significance level, there is no theoretical reason which prevents the use of other levels of confidence, for example, 90 or 99%. It should be clear that, if we want a 99% confidence interval for b instead of a 95% confidence interval, we are choosing to judge consistency on the basis of the 1% critical value for the distribution of T. According to table 8.1, this number is 2.576. If we replace 1.96, the 5% critical value, by 2.576, we will obtain $(\hat{b} - 2.576\hat{\sigma}, \hat{b} + 2.576\hat{\sigma})$, the 99% confidence interval for b. Or we could use $(\hat{b} - 1.645\hat{\sigma}, \hat{b} + 1.645\hat{\sigma})$, the 90% confidence interval for b.

From these formulae, we can make an important observation about the relationship between the length of the confidence interval and the level of confidence. The length of the 90% interval is $2(1.645)\hat{\sigma} = 3.29\hat{\sigma}$; the corresponding values for the 95 and 99% intervals are $3.92\hat{\sigma}$ and $5.152\hat{\sigma}$, respectively. Therefore, if we wish to increase our confidence that the interval includes the true value of b, the interval will necessarily be longer.

For the lymphoma data which we discussed in §8.3, we have stated that the method of chapter 13 leads to an estimate for b, the logarithm of the ratio of mortality rates, which is $\hat{b} = 1.10$; the estimated standard error of $\hat{b}$ is $\hat{\sigma} = 0.41$. Therefore, a 95% confidence interval for the true value of b is the set of values

$$\{1.10 - 1.96(0.41), 1.10 + 1.96(0.41)\} = (0.30, 1.90).$$

But what information is conveyed by this interval of values for b? In particular, does this 95% confidence interval provide information concerning the survival of lymphoma patients which we could not deduce from the significance test which was discussed in §8.3?

The test of significance prompted us to conclude that symptom classification does influence survival. Given this conclusion, a token examination of the data would indicate that the mortality rate is higher for those patients who present with B symptoms. Notice that this same information can also be deduced from the confidence interval. The value b = 0, which corresponds to equal survival experience in the two groups, is excluded from the 95% confidence interval for b; therefore, in light of the data, b = 0 is not a plausible value. Moreover, all values in the interval are positive, corresponding to a higher estimated mortality rate in the patients with B symptoms.

But there is additional practical information in the 95% confidence interval for b. The data generate an estimate of $\exp(\hat{b}) = 3.0$ for the ratio of the mortality rates. However, the width of the confidence interval for b characterizes the precision of this estimate. Although 3.0 is, in some sense, our best estimate of the ratio of the mortality rates, we know that the data are also consistent with a ratio as low as exp(0.30) = 1.35, or as high as exp(1.90) = 6.69. If knowledge that this ratio is at least as high as 1.35 is sufficient grounds to justify a change in treatment patterns, then such an alteration might be implemented. On the other hand, if a ratio of 2.0, for example, was necessary to justify the change, then the researchers would know that additional data were needed in order to make the estimation of b more precise. In general, a narrow confidence interval reflects the fact that we have fairly precise knowledge concerning the value of b, whereas a wide confidence interval indicates that our knowledge of b, and the effect it represents, is rather imprecise and therefore less informative.

In succeeding chapters, we will use significance tests, parameter estimates and confidence intervals to present the analyses of medical studies which demonstrate the use of more advanced statistical methodology. By emphasizing this combined approach to the analysis of data, we hope to make confidence intervals more attractive, and more familiar, to medical researchers.

It is probably true that most medical researchers will have little need to know more than the selected critical values presented in table 8.1. Nevertheless, precise p-values are often quoted in the medical literature, and we will do the same in subsequent chapters. The next section describes how to use a table of the cumulative probability curve for the standardized normal distribution. The exposition is necessarily somewhat detailed. Therefore, readers who are willing to accept, on trust, a few precise p-values may safely omit §8.5 and proceed to chapter 9.

8.5. Using Normal Tables

We have already discovered, in §8.2, that probabilities for $X \sim N(\mu, \sigma^2)$ can be evaluated by computing the equivalent probabilities for the standardized random variable $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Therefore, to calculate any probability for a normal distribution, we only require a table of values of the cumulative probability curve for $Z \sim N(0, 1)$. Such a table of values would still be fairly large if we neglect to exploit additional properties of the standardized normal distribution. Since the total amount of probability in the distribution of Z is always one, it follows that if z is any number we choose, $\Pr(Z > z) = 1 - \Pr(Z \leq z)$; therefore, if a table for the standardized normal distribution specifies the values of $\Pr(Z \leq z)$ as z varies, we can always calculate $\Pr(Z > z)$. More important, since the standardized normal distribution is symmetric about zero, we only need a table of values for $\Pr(Z \leq z)$ when $z \geq 0$ because

$$\Pr(Z < -z) = \Pr(Z > z) = 1 - \Pr(Z \leq z).$$

Table 8.2 specifies values of the cumulative probability curve, $\Pr(Z \leq z) = \Phi(z)$, for the standardized normal random variable Z. Here we have introduced the symbol $\Phi(z)$ to represent the value of the cumulative probability curve at z. The entry in the table at the intersection of the row labelled 2.5 and the column labelled 0.02 is the value of $\Phi(2.50 + 0.02) = \Phi(2.52) = \Pr(Z \leq 2.52) = 0.9941$. To calculate the probability that Z is at most 1.56, say, locate $\Pr(Z \leq 1.56) = \Phi(1.56)$ at the intersection of the row labelled 1.5 and the column labelled 0.06, since $1.56 = 1.50 + 0.06$; the value of $\Phi(1.56)$ is 0.9406. Similarly, $\Phi(0.74) = 0.7704$ and $\Phi(2.32) = 0.9898$. To evaluate $\Phi(-0.74)$ and $\Phi(-2.32)$, we must use the fact that if $z \geq 0$,

$$\Phi(-z) = \Pr(Z \leq -z) = \Pr(Z \geq z) = 1 - \Pr(Z < z) = 1 - \Phi(z).$$

Then, $\Phi(-0.74) = 1 - \Phi(0.74) = 1 - 0.7704 = 0.2296$ and $\Phi(-2.32) = 1 - \Phi(2.32) = 1 - 0.9898 = 0.0102$.

Table 8.2. Values of the cumulative probability function, $\Phi(z) = \Pr(Z \leq z)$, for the standardized normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Adapted from: Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol 1, pp 110–116. It appears here with the kind permission of the publishers.

To calculate $\Pr(a < Z < b)$, where a and b are two numbers, we proceed as follows:

$$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z \leq a) = \Phi(b) - \Phi(a).$$

Thus, if a = 1.0 and b = 2.5 we have

$$\Pr(1.0 < Z < 2.5) = \Phi(2.5) - \Phi(1.0) = 0.9938 - 0.8413 = 0.1525.$$

Similarly, when a = –1.0 and b = 2.5 we obtain

$$\begin{aligned}\Pr(-1.0 < Z < 2.5) &= \Phi(2.5) - \Phi(-1.0)\\ &= 0.9938 - \{1 - \Phi(1.0)\} = 0.9938 - 0.1587 = 0.8351.\end{aligned}$$

Thus far, we have only used table 8.2 to calculate probabilities for $Z \sim N(0, 1)$. To handle probabilities for $X \sim N(\mu, \sigma^2)$, we need to use the standardizing transformation, i.e., $\frac{X-\mu}{\sigma} = Z \sim N(0, 1)$. Provided we know the values of μ and σ, we can write

$$\Pr(X \leq x) = \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Pr\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

For example, if $X \sim N(4, 25)$, i.e., $\mu = 4$, $\sigma = 5$, then

$$\Pr(X \leq 4) = \Pr\left(Z \leq \frac{4-4}{5}\right) = \Phi(0) = 0.500.$$

Likewise,

$$\Pr(X \leq 2) = \Pr\left(Z \leq \frac{2-4}{5}\right) = \Phi(-0.40) = 1 - \Phi(0.40) = 0.3446.$$

As a final example in the use of normal tables, we evaluate $\Pr(X > 8.49)$ when $X \sim N(3.21, 3.58)$, i.e., $\mu = 3.21$, $\sigma = \sqrt{3.58} = 1.892$:

$$\Pr(X > 8.49) = \Pr\left(Z > \frac{8.49 - 3.21}{1.892}\right) = \Pr(Z > 2.79) = 1 - \Phi(2.79) = 0.0026.$$

Many software packages will evaluate probabilities from the normal distribution. However, to understand the output from a software package properly, it is useful to know how to evaluate a probability for $X \sim N(\mu, \sigma^2)$ using the equivalent probabilities for the standardized normal random variable $Z \sim N(0,1)$. Indeed, in some cases, a package may only avoid the final step in the process outlined here, which involves reading the numerical value from table 8.2. In this case, understanding how probabilities for the standardized normal random variable can be used more generally is essential.

9

Analyzing Normally Distributed Data

9.1. Introduction

The methods described in this chapter feature prominently in most introductory textbooks on statistics. In areas of research such as the physical sciences and engineering, measurements are frequently made on a continuous scale. The assumption that such data have a normal distribution has often proved to be a satisfactory description of the observed variation. Methodology for analyzing normally distributed data is therefore important, and deserves its widespread use.

The nature of the data collected in clinical research often precludes the use of these specialized techniques. Consequently, we have emphasized simple methods of analysis which are appropriate for the type of data frequently seen in medical research, especially that concerned with chronic disease. While it is true that none of the material in this chapter is critical to an understanding of the rest of the book, statistical techniques for normally distributed data can be quite useful, and we would be negligent if we failed to mention them altogether. Nonetheless, the discussion involves more formulae than we usually include in one chapter, and readers who find it too taxing should proceed to chapter 10.

Table 9.1 presents the results of an immunological assay carried out on 14 hemophiliacs and 33 normal controls; the test was performed at two concentrations, low and high. The primary purpose of the study was to ascertain if immunological differences between hemophiliacs and normal individuals could be detected. This is one type of data arising in medical research for which the use of statistical techniques for data from a normal distribution may be appropriate. Throughout this chapter, these data will be used to illustrate the methods of analysis which will be discussed.

Table 9.1. The results of an immunological assay of 33 normal controls and 14 hemophiliacs

Controls				Hemophiliacs	
concentration		concentration		concentration	
low	high	low	high	low	high
13.5	25.2	49.2	60.7	11.0	29.0
16.9	44.8	71.5	76.1	9.8	20.3
38.3	62.3	23.3	31.5	61.2	71.2
23.2	47.1	46.1	74.7	63.4	89.9
27.6	39.8	44.5	70.6	11.1	32.4
22.1	44.6	49.4	63.6	8.0	9.9
33.4	54.1	27.2	35.2	40.9	64.3
55.0	55.5	30.6	49.8	47.7	79.1
66.9	86.2	26.1	41.3	19.3	40.2
78.6	102.1	71.5	129.3	18.0	33.7
36.3	88.2	26.6	66.6	24.6	51.8
66.6	68.0	36.9	32.4	39.6	61.4
53.0	92.5	49.5	59.4	24.4	39.3
49.7	73.6	32.8	58.9	11.3	32.8
26.7	40.3	7.9	32.4		
62.9	93.8	9.6	30.0		
46.4	65.4				

9.2. Some Preliminary Considerations

9.2.1. Checking the Normal Assumption

A number of statisticians have investigated the robustness of methods for analyzing normal data. In general, the results of this research suggest that even if the distribution of the data is moderately non-normal, the use of specialized methods for normally distributed data is unlikely to seriously misrepresent the true situation. Nevertheless, it is patently careless to apply the techniques of §§9.3–9.5 to a set of observations without first checking that the assumption of a normal distribution for those observations is reasonable, or at least roughly true. Despite the appropriateness of these specialized methods in diverse circumstances, it is probably also true that the same techniques are frequently used in biological settings where the assumption of a normal distribution is not justified.

There are a number of ways in which the reasonableness of the normal distribution assumption may be verified. Perhaps the most straightforward ap-

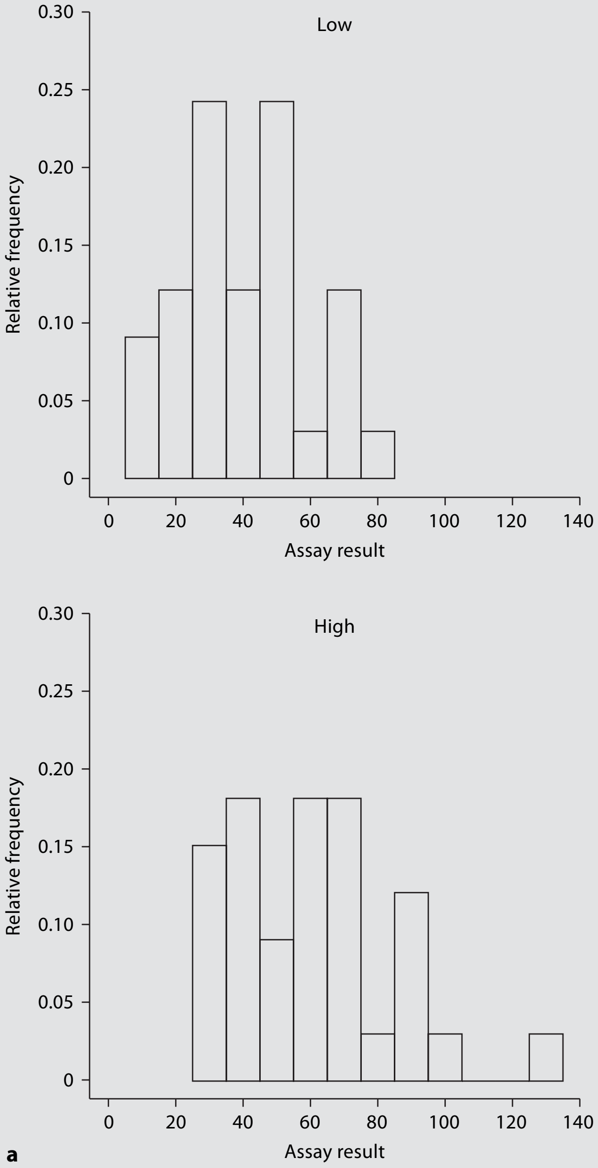


Fig. 9.1a. Histograms of the low and high concentration immunological assay results for the control sample of 33 individuals. Original data.

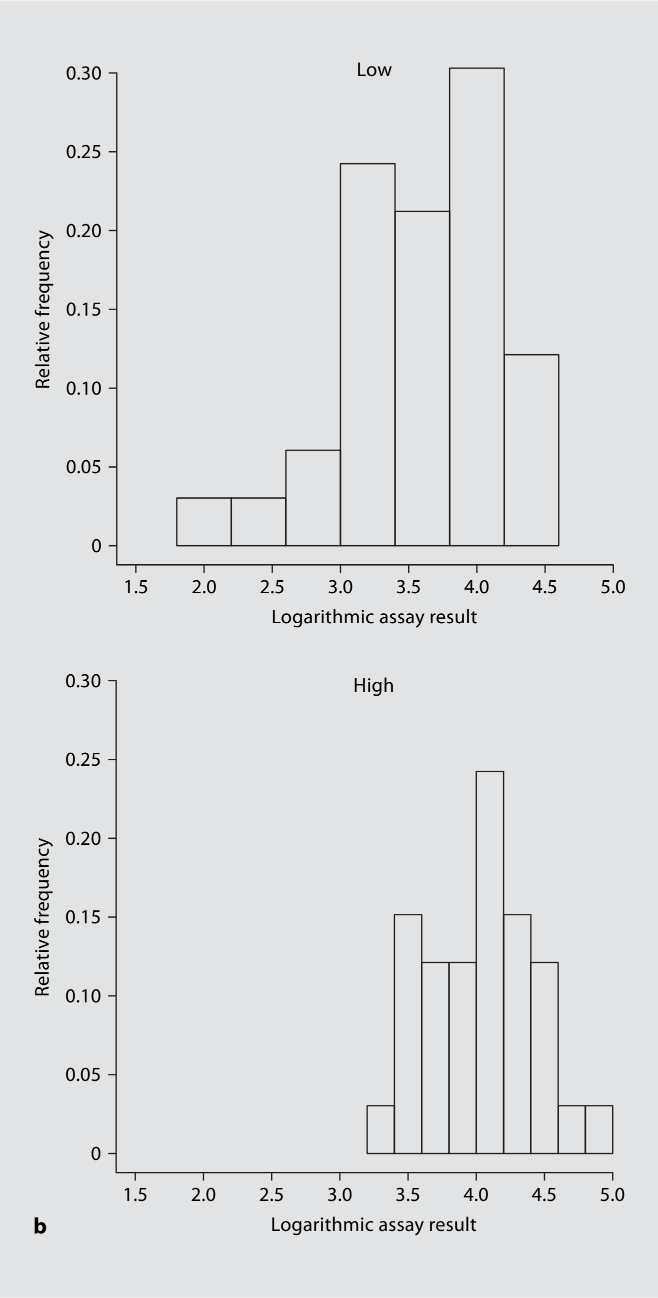


Fig. 9.1b. Logarithms of the original data.

proach involves comparing a histogram for the data with the shape of the normal probability curve. In §1.2, we briefly described how to construct a histogram for a set of data. More detailed instructions can always be found in any elementary text on statistics. Despite the degree of subjectivity which is involved in constructing a histogram, e.g., in the choice of intervals, etc., it is wise to conclude that if the histogram does not appear roughly normal, then the use of methods which are described in this chapter may not be appropriate.

Figure 9.1a presents histograms of the high and low concentration assay results for the control samples, cf. table 9.1. We can see that the shapes of these two histograms are not radically different from the characteristic shape of a normal probability curve. However, a few large assay values give the histograms a certain asymmetry which is not characteristic of the normal distribution. To correct this problem, one might consider transforming the data. For example, figure 9.1b shows histograms for the logarithms of the high and low concentration assay results. On a log scale, the large values are perhaps less extreme, but the shapes of the corresponding histograms are not much closer to the characteristic shape of the normal distribution than those of the original data. Although other transformations could be tried, the histograms in figure 9.1a are not sufficiently non-normal to preclude at least a tentative assumption that the data are normally distributed. Therefore, we will proceed from this assumption and regard the measured values as normal observations. In §9.4.1, the need to use a transformation will be more apparent.

9.2.2. Estimating the Mean and Variance

Except in unusual circumstances where relevant prior information is available concerning the population, the mean, μ, and the variance, σ^2, will need to be estimated from the data. A natural estimate of the population mean is the sample mean. Symbolically, if x_i represents the i^{th} observation in a sample of n values, so that $x_1, x_2, ..., x_n$ represent the entire sample, then the formula for $\hat{\mu}$, the estimator of the population mean μ, is

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

In §1.3, we described the variance of a distribution as the expected value of the constructed random variable $\{X - E(X)\}^2 = (X - \mu)^2$. If we knew the value of μ, the mean of the distribution, a natural estimate of the population variance, σ^2, would be

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}\{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2\}.$$

Table 9.2. Sample means and sample standard deviations for the immunological assay results presented in table 9.1

	Low concentration		High concentration	
	controls	hemophiliacs	controls	hemophiliacs
n	33	14	33	14
$\Sigma\, x_i$	1319.8	390.3	1996.0	655.3
$\Sigma\, x_i^2$	64212.98	15710.21	138906.30	37768.87
$\overline{x}$	39.99	27.88	60.48	46.81
s^2	357.16	371.48	568.08	545.86
s	18.90	19.27	23.83	23.36

Example calculation: controls, low concentration

$\overline{x} = \frac{1319.8}{33} = 39.99$, $s^2 = \frac{1}{32}\{64212.98 - 33(39.99)^2\} = 357.16$, $s = \sqrt{357.16} = 18.90$.

Since we do not know the value of μ, we can substitute $\hat{\mu} = \overline{x}$ in the above formula. We should realize that this substitution introduces additional uncertainty into our estimate of σ^2, since $\hat{\mu}$ is only an estimate of μ. For reasons which we cannot explain here, the use of $\overline{x}$ also leads to the replacement of n by (n – 1) in the divisor. The formula for s^2, the estimator of the population variance, σ^2, then becomes

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\}.$$

After a certain amount of algebraic manipulation, this formula can be rewritten as

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right\}.$$

Both formulae are correct; however, the second expression simplifies the amount of calculation considerably. To estimate the standard deviation of the data we simply use s, the positive square root of s^2.

From the statistical point of view, $\overline{x}$ and s^2 are, in some sense, the best estimates of μ and σ^2 that we can obtain from normally distributed data, i.e., from $x_1, x_2, \ldots, x_n$. However, it is wise to remember that for other distributions, $\overline{x}$ and s^2 may not necessarily be the best estimates of μ and σ^2 in the same sense.

Table 9.2 shows the detailed calculations which lead to the estimates $\overline{x}$ and s for the low and high concentration assay results presented in table 9.1. The values of $\overline{x}$ and s could be used to estimate a plausible range of values for the immunological test procedure; however, this information does not explicitly address the question of immunological differences between the hemophiliac and control populations. To investigate this question, we require the methods which are described in §9.4.

9.3. Analyzing a Single Sample

In most practical circumstances involving a single sample from a particular population, the principal question which is of interest concerns the value of μ, the population mean. Either we may wish to test the hypothesis that μ is equal to a specific value μ_0, say, or we may wish to derive a confidence interval for μ. We know that the sample average is a natural estimate of μ. Therefore, it should not be surprising that the sample average is used to test the hypothesis H: $\mu = \mu_0$, and also to derive a suitable confidence interval for μ.

If we let the random variables $X_1, X_2, \ldots, X_n$ represent a single sample of n observations, then the assumption that these data are normally distributed can be specified, symbolically, by writing $X_i \sim N(\mu, \sigma^2)$ for i = 1, 2, …, n. It can also be shown that the sample average, $\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$, has a normal distribution with mean μ and variance σ^2/n, i.e., $\overline{X} \sim N(\mu, \sigma^2/n)$. Therefore, the standardizing transformation of chapter 8 – which represents the distance between $\overline{X}$ and μ in units of the standard error of $\overline{X}$ – guarantees that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

and a suitable test statistic for evaluating the significance level of the data with respect to the hypothesis H: $\mu = \mu_0$ is

$$T = \frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{n}}.$$

If σ is known and t_o is the observed value of T, then the significance level of the test is $\Pr(T \geq t_o)$. As we discovered in §8.4, the calculations which generate a confidence interval for μ are based on the fact that $T = |Z|$, where $Z \sim N(0, 1)$. The formula specifying a 95% confidence interval for μ is the interval

$$(\overline{x} - 1.96\sigma/\sqrt{n}, \overline{x} + 1.96\sigma/\sqrt{n}),$$

where $\overline{x}$ is the observed sample mean.

In most cases, σ will not be known. An obvious solution to this problem involves replacing σ by its estimate, s. In this case, the significance level is only approximate, since $Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ no longer has a normal distribution; we have estimated σ by s. However, if the number of observations is large, i.e., $n \geq 50$,

Table 9.3. Approximate 95% confidence intervals for the mean immunological assay results, at low and high concentrations, in the hemophiliac and control populations

	Concentration	
	low	high
Controls	(33.54, 46.44)	(52.35, 68.61)
Hemophiliacs	(17.79, 37.97)	(34.57, 59.05)

Example calculation: controls, low concentration (n = 33)

$\overline{x} = 39.99$ $\quad 1.96s/\sqrt{n} = \dfrac{1.96(18.90)}{\sqrt{33}} = 6.45$

$s = 18.90$ $\quad \overline{x} - 6.45 = 33.54, \overline{x} + 6.45 = 46.44.$

the approximation will be quite accurate. Similar comments apply to the approximate 95% confidence interval ($\overline{x} - 1.96s/\sqrt{n}, \overline{x} + 1.96s/\sqrt{n}$).

If we apply these methods to the immunological data discussed in §§9.1, 9.2, we obtain two approximate 95% confidence intervals for the mean assay results in each of the study groups. The results of these calculations are given in table 9.3. Since there is no previous information regarding these assays in either the hemophiliacs or the controls, there is no natural hypothesis concerning μ, i.e., no obvious value, μ_0, that we might consider testing with these data.

In the preceding discussion, we remarked that when σ is replaced by s in the formula for T and the corresponding 95% confidence interval for μ, the results we obtain are only approximate. This is because s is an estimate of σ. As a result, we are more uncertain about the precise value of μ. In particular, if the sample size is rather small, e.g., $n < 30$, say, we ought to use the exact distribution of $T = \frac{|\overline{X} - \mu_0|}{s/\sqrt{n}}$ to calculate the significance level of the data with respect to the hypothesis H: $\mu = \mu_0$, and also to obtain a 95% confidence interval for μ. Now, if the null hypothesis is true, it can be shown that the statistic $\frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ has a Student's t distribution on (n – 1) degrees of freedom; symbolically, we write this as $\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$. The probability curve for the Student's t distribution is similar to that of the standardized normal distribution. Both curves are symmetric about zero and are approximately bell-shaped; however, the curve for $t_{(n-1)}$ is somewhat more spread out than the probability curve of Z. In fact, the spread of the Student's t distribution depends on a parameter called the degrees of freedom. Since this parameter is equal to (n – 1), i.e., sample size minus one, the degrees of freedom reflect the number of observations used to estimate σ and, therefore, how accurate an estimate s is likely to be. A Student's t distribution with large degrees of freedom, say 50, is virtually indistinguishable from a standardized normal distribution.

Statistical tables for the Student's t distribution are usually drawn up on the same principle as a table of critical values for the distribution of $|Z|$ (cf., table 8.1). Each row of the table corresponds to an integer value of the degrees of freedom, and each column of the table represents a specified probability level; thus, the values in the body of the table are critical values for the distribution of $T = |t_{(k)}|$. For an example of Student's t distribution tables which correspond to this format, see table 9.4. Critical values for the distribution of $|Z|$ may be found in the last row of table 9.4, since a $t_{(\infty)}$ distribution is identical with the distribution of Z. Notice that these values are all smaller than the corresponding critical values found in other rows of the table; this difference reflects the fact that the spread of the Student's t distribution is greater than that of the standardized normal distribution.

A suitable test statistic for evaluating the significance level of the data with respect to the hypothesis H: $\mu = \mu_0$ is

$$T = \frac{|\overline{X} - \mu_0|}{s/\sqrt{n}};$$

since $\frac{\overline{X}-\mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$ if the null hypothesis is true, it follows that $T \sim |t_{(n-1)}|$. Therefore, if t_o is the observed value of T, the significance level is equal to

$$\Pr(T \geq t_o) = \Pr(|t_{(n-1)}| \geq t_o);$$

this is the reason that table 9.4 presents critical values of the distribution of $T = |t_{(k)}|$. Similar calculations lead to the formula

$$(\bar{x} - t^* s/\sqrt{n}, \bar{x} + t^* s/\sqrt{n}),$$

which specifies a 95% confidence interval for μ, where t^* is the appropriate 5% critical value from table 9.4, i.e., $\Pr(T \geq t^*) = \Pr(|t_{(n-1)}| \geq t^*) = 0.05$. Notice that this version of the 95% confidence interval for μ differs from the approximate confidence interval which we obtained at the beginning of this section only in the replacement of 1.96, the 5% critical value for $|Z|$, by t^*, the corresponding critical value for the distribution of $T = |t_{(n-1)}|$. Replacing 1.96 by t^* always increases the width of the confidence interval, since t^* exceeds 1.96. The increased width reflects our increased uncertainty concerning plausible values of μ since we have used s, an estimate of σ, to derive the 95% confidence interval.

Exact 95% confidence intervals for the mean assay results at low and high concentrations in the hemophiliac and control populations may be found in table 9.5; the table also shows selected details of the calculations. If we compare corresponding intervals in tables 9.3 and 9.5, we see that the exact 95% confi-

Table 9.4. Critical values of the probability distribution of $T = |t_{(k)}|$; the table specifies values of the number t_o such that $\Pr(T > t_o) = p$

Degrees of freedom (k)	Probability level, p					
	0.10	0.05	0.02	0.01	0.002	0.001
1	6.314	12.706	31.82	63.66	318.3	636.6
2	2.920	4.303	6.695	9.925	22.33	31.60
3	2.353	3.182	4.541	5.841	10.21	12.92
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.767
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646
40	1.684	2.021	2.423	2.704	3.307	3.551
60	1.671	2.000	2.390	2.660	3.232	3.460
120	1.658	1.980	2.358	2.617	3.160	3.373
∞ (normal)	1.645	1.960	2.326	2.576	3.090	3.291

Abridged from Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol 1, p 146. It appears here with the kind permission of the publishers.

Table 9.5. Exact 95% confidence intervals for the mean immunological assay results, at low and high concentrations, in the hemophiliac and control populations

	Concentration	
	low	high
Controls	(33.27, 46.71)	(52.01, 68.95)
Hemophiliacs	(16.76, 39.00)	(33.32, 60.30)
Example calculations:	low concentration	
Hemophiliacs (n = 14):		
$\bar{x} = 27.88$ $s = 19.27$	$2.160s/\sqrt{n} = \frac{2.160(19.27)}{\sqrt{14}} = 11.12$[a] $\bar{x} - 11.12 = 16.76, \bar{x} + 11.12 = 39.00$	
Controls (n = 33):		
$\bar{x} = 39.99$ $s = 18.90$	$2.042s/\sqrt{n} = \frac{2.042(18.90)}{\sqrt{33}} = 6.72$[b] $\bar{x} - 6.72 = 33.27, \bar{x} + 6.72 = 46.71$	

[a] The 5% critical value for $|t_{(13)}|$ is 2.160 (see table 9.4).

[b] Since the 5% critical value for $|t_{(32)}|$ is not given in table 9.4, we use the corresponding value for $|t_{(30)}|$, which is 2.042.

dence interval is always wider than the corresponding approximate confidence interval. Since the separate 95% confidence intervals for hemophiliacs and controls overlap at both the low and high concentrations, we might reasonably conclude that, at each concentration level, the data do not contradict the hypothesis that the mean assay in the two populations is the same. An alternative method for investigating this question, which is essentially the problem of comparing two means in normally distributed data, is discussed in §9.4.

9.4. Comparisons Based on the Normal Distribution

9.4.1. Paired Data

In previous chapters, we discussed the importance of stratifying data in order to properly evaluate comparisons that were of interest. The basic premise of stratification is that any comparison with respect to a particular factor should be made between groups which are alike with respect to other factors that may influence the response. Data which are naturally paired constitute a special case of stratification, and typically facilitate a more precise comparison than might otherwise be achieved. Before and after measurements are perhaps the most common illustration of naturally paired data; for example, we might wish to investigate the efficacy of an anti-hypertensive drug by using measurements of a patient's blood pressure before and after the drug is administered. In another circumstance, if we had identical twins who differed with respect to smoking status, we might consider measuring several aspects of lung function in each twin in order to investigate the effect of smoking on lung function.

When data involve natural pairing, a common method of analysis is based on the assumption that the differences between the paired observations are normally distributed. If we let the random variables $D_1, D_2, \ldots, D_n$ represent these differences, e.g., Before-After, then we are assuming that $D_i \sim N(\mu_d, \sigma_d^2)$ for $i = 1, 2, \ldots, n$; μ_d is the mean difference in response, and σ_d^2 is the variance of the differences. Provided this assumption is reasonable, then all the methods for a single sample of normally distributed data can be used to analyze the differences. For example, a natural hypothesis to test in this situation would be H: $\mu_d = 0$; i.e., the mean difference in response is zero. If the data provide evidence to contradict this hypothesis, then we would conclude that the factor which varies within each pair, e.g., smoking status, or use of the anti-hypertensive drug, has a real effect. We might also want to evaluate a suitable confidence interval for μ_d in order to estimate the magnitude and direction of the effect which has been detected.

The immunological data which we have previously discussed do involve natural pairing since, for each subject, we have one assay result at each concentration level. If we compute the High-Low differences of the assay results for each patient, we could separately investigate the effect of concentration among hemophiliacs, and also among the controls. Although this question is not of primary interest to the study, we shall treat the data for the controls as we have described in order to illustrate the analysis of paired observations.

The original assay results may be found in table 9.1. The difference, High – Low, was calculated for each control subject, yielding 33 observations. A histogram of this sample of differences is given in figure 9.2a. The distribution of the differences is quite spread out, and one might be reluctant to assume that the distribution is normal. An alternative approach would be to consider the differences in the logarithms of the assay results. This is equivalent to considering the logarithm of the ratio of the assay results. A histogram of the logarithm differences is given in figure 9.2b, and it can be seen that the shape of this histogram is considerably closer to the characteristic shape of a normal probability curve than that of the histogram in figure 9.2a. Therefore, we will analyze the differences between the logarithms of the assay results at the two concentrations, and assume that these differences are normally distributed.

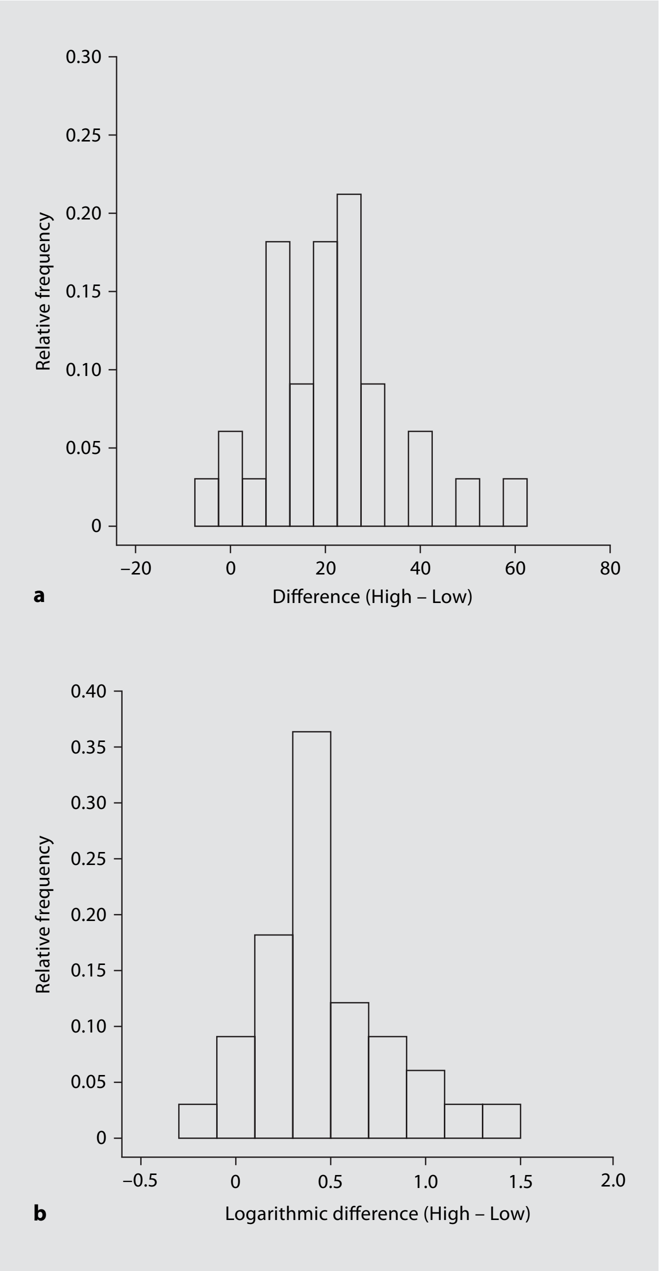


Fig. 9.2. Histograms of the difference in immunological assay results at high and low concentrations for the control sample of 33 individuals. **a** Original data. **b** Logarithms of the original data.

For these logarithm differences we obtain $\overline{d} = 0.47$ and $s = 0.33$. The observed value of T, the statistic for testing H: $\mu_d = 0$, is

$$t_0 = \frac{|\overline{d} - 0|}{s/\sqrt{n}} = \frac{0.47}{0.33/\sqrt{33}} = 8.18.$$

As we might expect, this large observed value tells us that the data provide strong evidence to contradict the hypothesis that the mean logarithm difference is zero ($p < 0.001$); we therefore conclude that concentration influences the assay result.

The magnitude of this effect, i.e., the change with concentration, is indicated by the estimated mean of 0.47, and also by the 95% confidence interval for μ_d, which is (0.35, 0.59). In terms of the original assay measurements obtained on each control subject at the Low and High concentration levels, these results represent a corresponding estimate for the ratio of High to Low concentration assay results of $e^{0.47} = 1.6$; the corresponding 95% confidence interval is $(e^{0.35}, e^{0.59}) = (1.42, 1.80)$.

In many circumstances, it is not possible to make comparisons which are based on naturally paired data. Methods which are appropriate for this more common situation are discussed in the next section.

9.4.2. Unpaired Data

If the data are not naturally paired, then the comparison of interest usually involves two separate, independent samples from two (assumed) normal distributions. We can portray this situation, symbolically, by using the random variables $X_1, X_2, \ldots, X_n$ to represent one sample and $Y_1, Y_2, \ldots, Y_m$ to represent the second sample; then $X_i \sim N(\mu_x, \sigma^2)$ for $i = 1, 2, \ldots, n$ and $Y_j \sim N(\mu_y, \sigma^2)$ for $j = 1, 2, \ldots, m$. For example, the X's may be the results of lung function tests for a random sample of smokers, and the Y's may be a corresponding set of observations on nonsmokers. Notice that whereas the means, μ_x and μ_y, of the two distributions may differ, the corresponding variances are equal to the same value, σ^2. The primary question of interest in such a situation is usually a comparison of the means, μ_x and μ_y.

Since $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_{i,}$ and $\overline{Y} = \frac{1}{m} \sum_{j=1}^{m} Y_{j,}$ are natural estimates of μ_x and μ_y, the obvious quantity on which to base a comparison of μ_x and μ_y is $\overline{X} - \overline{Y}$, the difference in sample means. Conveniently, $\overline{X} - \overline{Y}$ has a normal distribution with mean $(\mu_x - \mu_y)$ and variance $\sigma^2(1/n + 1/m)$. When σ is known, a confidence interval for $(\mu_x - \mu_y)$ or a significance test concerning this difference can be based on the normal distribution of $\overline{X} - \overline{Y}$. In most cases, the value of σ will not be known. However, if we replace σ by a suitable estimate, s, then it can be shown that

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{s\sqrt{1/n + 1/m}} \sim t_{(n+m-2)}.$$

In view of the results which we described in §9.3, this is the distribution that we would expect to obtain after replacing σ by s. To estimate σ^2, the common variance in the two populations, we use a weighted average of

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right\}$$

and

$$s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_j - \overline{y})^2 = \frac{1}{m-1}\left\{\sum_{j=1}^{m} y_j^2 - m\overline{y}^2\right\},$$

the individual sample estimates. The formula for the combined or pooled estimate of σ^2 is

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

To test the hypothesis that $(\mu_x - \mu_y)$ is equal to δ_0, say, we would use the test statistic

$$T = \frac{|\overline{X} - \overline{Y} - \delta_0|}{s\sqrt{1/n + 1/m}}.$$

In many circumstances it is the hypothesis of equal means, H: $\mu_x = \mu_y$, i.e., $\mu_x - \mu_y = \delta_0 = 0$, which is of primary interest. If the null hypothesis is true, $T \sim |t_{(n+m-2)}|$. Therefore, if t_o is the observed value of T, the significance level of the data is equal to $\Pr(T \geq t_o) = \Pr(|t_{(n+m-2)}| \geq t_o)$; this value can be estimated using table 9.4.

If a 95% confidence interval for $(\mu_x - \mu_y)$ is required, the formula which is obtained from the usual calculations is

$$(\overline{x} - \overline{y} - t^* s\sqrt{1/n + 1/m}, \overline{x} - \overline{y} + t^* s\sqrt{1/n + 1/m}),$$

where $\overline{x}$ and $\overline{y}$ are the observed sample means, and t^* is the appropriate 5% critical value from table 9.4, i.e., $\Pr(T \geq t^*) = \Pr(|t_{(n+m-2)}| \geq t^*) = 0.05$.

The purpose of the immunological study was to characterize differences between the hemophiliacs and the controls. The techniques for unpaired data which we have just described allow us to address this question explicitly. We shall look, separately, at the two concentration levels, using X's to represent the assay results for hemophiliacs and Y's for the results in the controls; summary statistics for each group may be found in table 9.6. A natural hypothesis to test in this data set is H: $\mu_x = \mu_y$, i.e., $\mu_x - \mu_y = 0$. Details of the actual calculations which are required to test this hypothesis, at each concentration, may be found in table 9.6.

Table 9.6. Comparing the mean immunological assay results, at low and high concentrations, in the hemophiliac and control populations

	Concentration	
	low	high
Hemophiliacs (n = 14)	$\bar{x} = 27.88$, $s_x^2 = 371.48$	$\bar{x} = 46.81$, $s_x^2 = 545.86$
Controls (m = 33)	$\bar{y} = 39.99$, $s_y^2 = 357.16$	$\bar{y} = 60.48$, $s_y^2 = 568.08$
Pooled estimate, $s^2 = \frac{13s_x^2 + 32s_y^2}{45}$	$\frac{13(371.48) + 32(357.16)}{45} = 361.30$	$\frac{13(545.86) + 32(568.08)}{45} = 561.66$
Observed value, $t_o = \frac{\lvert \bar{x} - \bar{y} \rvert}{s\sqrt{1/14 + 1/33}}$	$\frac{\lvert 27.88 - 39.99 \rvert}{19.01\sqrt{1/14 + 1/33}} = 2.00$	$\frac{\lvert 46.81 - 60.48 \rvert}{23.70\sqrt{1/14 + 1/33}} = 1.81$
Significance level, $\Pr(\lvert t_{(45)} \rvert \geq t_o)$	0.05	0.05 – 0.10
95% confidence interval for $(\mu_x - \mu_y)$, $(\bar{x} - \bar{y}) \pm 2.021\, s\sqrt{1/14 + 1/33}$[a]	(–24.36, 0.14)	(–28.95, 1.61)

[a] Since the 5% critical value for $\lvert t_{(45)} \rvert$ is not given in table 9.4, we use the corresponding value for $\lvert t_{(40)} \rvert$, which is 2.021.

The significance level for a test of the hypothesis that $\mu_x = \mu_y$ is $\Pr(\lvert t_{(45)} \rvert \geq t_o)$, where t_o is the observed value of the test statistic. Since there is no row in table 9.4 corresponding to 45 degrees of freedom, we compare t_o with the 5% critical values for both 40 and 60 degrees of freedom. These values are 2.021 and 2.00, respectively. For the low concentration results, the value of t_o is 2.00, which corresponds, approximately, to a significance level of 0.05. The observed value of the test statistic for the high concentration assays is 1.81, which lies between the 5 and 10% critical values for both 40 and 60 degrees of freedom; thus, the p-value associated with this test of significance is between 0.05 and 0.10. From these results we conclude that there is suggestive evidence, at both low and high concentrations, for different mean assay results in the two populations. The 95% confidence intervals for $(\mu_x - \mu_y)$ which are given in table 9.6 include zero near the upper endpoint of each interval, and are consistent with this conclusion concerning μ_x and μ_y.

Notice that the evidence concerning different population mean values is somewhat stronger in table 9.6 than that which might be derived from an informal comparison of the confidence intervals presented in table 9.5. This is largely because the scale factor $\sqrt{1/n + 1/m}$ in the formula for the 95% confidence interval for $(\mu_x - \mu_y)$ is considerably smaller than the sum of the scale factors $\sqrt{1/n}$ and $\sqrt{1/m}$ that appear in corresponding formulae for the separate 95% confidence intervals for μ_x and μ_y. A method for examining the validity of the pooled estimate is described in the next section.

9.5. Testing the Equality of Variances

A critical assumption in the analysis of unpaired data which we described in §9.4.2 is the requirement that the variance in each population is the common value σ^2. It is difficult to imagine a situation where this requirement could be assumed to hold without checking its reasonableness. Let us suppose that $X_1, X_2, ..., X_n$ and $Y_1, Y_2, \ldots, Y_m$ represent unpaired samples of normally distributed observations. If s_x^2 and s_y^2 are the sample estimates of the variances σ_x^2 and σ_y^2, then the appropriate statistic for testing the hypothesis that the variances are equal, i.e., H: $\sigma_x^2 = \sigma_y^2 = \sigma^2$, is the ratio

$$R = \frac{s_x^2}{s_y^2},$$

where s_x^2 is assumed to be greater than s_y^2. If s_y^2 exceeds s_x^2, then the roles of the two samples can simply be interchanged.

If the null hypothesis that $\sigma_x^2 = \sigma_y^2 = \sigma^2$ is true, the ratio R has a probability distribution which depends on the F-distribution. This latter distribution is characterized by two parameters, the degrees of freedom associated with s_x^2, the greater variance estimate, and the degrees of freedom associated with s_y^2, the lesser variance estimate. Since the values of s_x^2 and s_y^2 were calculated from n and m observations, respectively, the corresponding degrees of freedom are (n – 1) and (m – 1). Therefore, if the null hypothesis is true, and if r_o is the observed value of R, the significance level of the data is equal to

$$\Pr(R \geq r_o) = 2\Pr(F_{n-1,\, m-1} \geq r_o).$$

Table 9.7 gives selected critical values for a number of different F-distributions. In using table 9.7, or any of the more extensive sets of statistical tables for the F-distribution which are available, an observed value of the ratio R should be compared with the entry in the statistical table which has degrees of freedom closest to (n – 1) and (m – 1), if an exact match is not possible. The use of more extensive tables will usually permit a more precise determination of the significance level. If the observed value of R is at all large, the appropriateness of any method for analyzing normal data which is based on the common variance assumption is questionable.

In the case of the immunological data, the observed values of R for the high and low concentration assay results can be calculated from the informa-

Table 9.7a. Selected 5% critical values of the F-distribution; the table gives values of the number r_o such that $Pr(F_{n,m} \geq r_o) = 0.05$

Degrees of freedom for the smaller variance estimate (m)	Degrees of freedom for the greater variance estimate (n)							
	1	2	3	4	5	6	12	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	243.9	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.41	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.74	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	5.91	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.68	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.00	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.57	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.28	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.07	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	2.91	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.79	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.69	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.60	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.53	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.48	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.42	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.38	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.34	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.31	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.28	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.25	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.23	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.20	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.18	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.16	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.15	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.13	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.12	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.10	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.09	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.00	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	1.92	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	1.83	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.75	1.00

Abridged from Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol 1, p 171. It appears here with the kind permission of the publishers.

Table 9.7b. Selected 2.5% critical values of the F-distribution; the table gives values of the number r_o such that $\Pr(F_{n,m} \geq r_o) = 0.025$

Degrees of freedom for the smaller variance estimate (m)	Degrees of freedom for the greater variance estimate (n)							
	1	2	3	4	5	6	12	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	976.7	1,018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.41	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.34	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	8.75	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.52	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.37	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.67	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.20	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	3.87	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.62	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.43	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.28	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.15	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.05	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	2.96	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	2.89	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	2.82	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	2.77	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	2.72	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	2.68	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.64	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.60	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.57	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.54	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.51	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.49	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.47	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.45	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.43	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.41	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.29	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.17	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.05	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	1.94	1.00

Abridged from Pearson ES, Hartley HO: Biometrika Tables for Statisticians. London, Biometrika Trustees, Cambridge University Press, 1954, vol 1, p 172. It appears here with the kind permission of the publishers.

Table 9.8. Testing the assumption of equal variances for the distribution of immunological assay results, at low and high concentrations, in the hemophiliac and control populations

Low concentration

Hemophiliacs:	$n = 14$	$s_x^2 = 371.48$
Controls:	$m = 33$	$s_y^2 = 357.16$

$$r_o = \frac{s_x^2}{s_y^2} = \frac{371.48}{357.16} = 1.04$$

$$\Pr(R \geq r_o) = 2\Pr(F_{13,32} \geq 1.04) > 0.10^a$$

High concentration

Controls:	$n = 33$	$s_x^2 = 568.08$	$r_o = 1.04$
Hemophiliacs:	$m = l4$	$s_y^2 = 545.86$	

$$\Pr(R \geq r_o) = 2\Pr(F_{32,13} \geq 1.04) > 0.10^b$$

[a] The 5% critical values for $F_{12,30}$ and $F_{12,40}$ are 2.09 and 2.00, respectively.
[b] The 5% critical values for $F_{12,13}$ and $F_{\infty,13}$ are 2.60 and 2.21, respectively.

tion given in table 9.6. Details of the calculations which are involved in testing the common variance assumption are presented in table 9.8. The equal variances assumption is not contradicted by the data at either concentration level.

If the data suggest that the variances in two unpaired samples are apparently different, then it is important to ask whether, in light of this information, a comparison of means is appropriate. A substantial difference in the variances may, in fact, be the most important finding concerning the data. Also, when the variances are different, the difference in means no longer adequately summarizes how the two groups differ.

If, after careful consideration, a comparison of means is still thought to be important, then the appropriate procedure is not uniformly agreed upon among statisticians. The reasons for this disagreement are beyond the scope of this book, but approximate methods should be adequate in most situations. Therefore, we suggest the simple approach of calculating a confidence interval for μ_x and a second interval for μ_y. If these intervals do not overlap, then the hypothesis that μ_x is equal to μ_y would also be contradicted by the data. Substantial overlap suggests consistency with the null hypothesis while minimal overlap corresponds to a problematic situation. In this latter case, or if the nature of the practical problem makes this ad hoc approach inappropriate or inadequate, a statistician should be consulted.

In chapter 10, we will also be discussing a method for analyzing normally distributed data. In fact, the techniques which we have described in chapter 9 can be regarded as a special case of the methodology which is presented in chapter 10. We do not intend to elaborate further on this connection, since the purposes of these two chapters are quite different. Chapter 9 constitutes a brief introduction to material which, although important, is not a primary focus of this book. However, in chapter 10 we introduce a class of statistical techniques which will feature prominently in the remaining chapters. We therefore direct our attention, at this point, to the important topic of regression models.

10

Linear Regression Models for Medical Data

10.1. Introduction

Many medical studies investigate the association of a number of different factors with an outcome of interest. In some of the previous chapters, we discussed methods of studying the role of a single factor, perhaps with stratification according to other factors to account for heterogeneity in the study population. When there is simultaneous interest in more than one factor, these techniques have limited application.

Regression models are frequently used in such multi-factor situations. These models take a variety of forms, but their common aim is to study the joint effect of a number of different factors on an outcome variable. The quantitative nature of the outcome variable often determines the particular choice of regression model. One type of model will be discussed in this chapter. The five succeeding chapters will introduce other regression models. In all six chapters, we intend to emphasize the general nature of these models and the types of questions which they are designed to answer. We hope that the usefulness of regression models will become apparent as our discussion of them proceeds.

Before we launch into a description of linear regression models and how they are used in medical statistics, it is probably important to make a few brief comments about the concept of a statistical model. A relationship such as Einstein's famous equation linking energy and mass, $E = mc^2$, is an exact and true description of the nature of things. Statistical models, however, use equations quite differently. The equation for a statistical model is not expected to be exactly true; instead, it represents a useful framework within which the statistician is able to study relationships which are of interest, such as the association between survival and age, aggressiveness of disease and the treatment which a patient receives. There is frequently no particular biological support for statis-

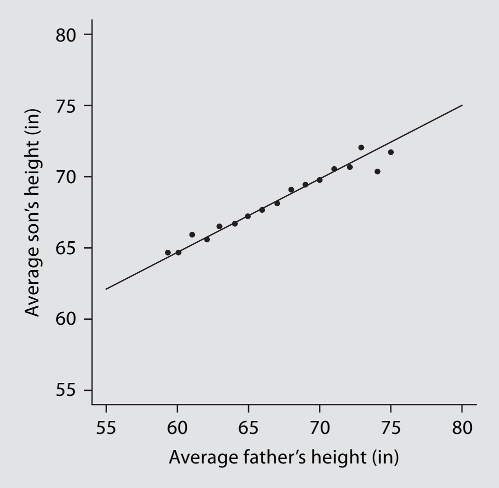


Fig. 10.1. A scatterplot of the data collected by Pearson and Lee [17] concerning the heights of father-son pairs. The equation of the fitted regression line is $\hat{Y} = 33.73 + 0.516X$.

tical models. In general, they should be regarded simply as an attempt to provide an empirical summary of observed data. Finally, no model can be routinely used without checking that it does indeed provide a reasonable description of the available data.

10.2. A Historical Note

The term 'regression' arose in the context of studying the heights of members of family groups. Pearson and Lee [17] collected data on the heights of 1078 father-son pairs in order to study Galton's 'law of universal regression' which states that 'Each peculiarity in a man is shared by his kinsmen, but *on the average* in a less degree'. Figure 10.1 shows a plot of the average height of the sons in each of 17 groups which were defined by first classifying the fathers into 17 groups, using one-inch intervals of height between 58.5 and 75.5 inches.

If we represent the height of a son by the random variable Y, and the height of a father by the random variable X, then the straight line drawn in figure 10.1 is simply the equation

$$\text{Son's height} = 33.73 + 0.516 \times \text{Father's height}$$

or

$$Y = 33.73 + 0.516\ X. \tag{10.1}$$

We shall use numbers in parentheses to label equations to which we later wish to refer. Equation (10.1) is called a regression equation. The variable Y is called the dependent, outcome or response variable, and X is called the independent or explanatory variable. Since the adjective independent is somewhat misleading, although widely used, we shall use the adjective explanatory. Another common term which we shall also use is covariate.

Obviously, not all the father-son pairs of heights lie along the drawn straight line. Equation (10.1) represents the best-fitting straight line of the general form

$$Y = a + bX, \tag{10.2}$$

where best-fitting is defined as the unique line which minimizes the average of the squared distances from each observed son's height to the chosen line. More specifically, we write the equation of the best-fitting line as

$$\hat{Y} = \hat{a} + \hat{b}X,$$

where $\hat{a}$ and $\hat{b}$ are best choices or estimates for a and b, and $\hat{Y}$ is the value of Y predicted by this model for a specified value of X. In general, statisticians use the ˆ notation to indicate that something is an estimate. The line in figure 10.1 minimizes the average of the values $(Y - \hat{Y})^2$. It is called a linear regression because Y is a straight-line function of the unknown parameters a and b.

Notice that equation (10.1) is an empirical relation, and that it does not imply a causal connection between X and Y.

In this example, the estimated line shows that there is a *regression* of sons' heights towards the average. This is indicated by the fact that $\hat{b}$, the multiplier of the fathers' heights, is much less than one. This historical terminology has persisted, so that an equation like (10.2) is still called a regression equation, and b is called a *regression coefficient,* whatever its value.

10.3. Multiple Linear Regression

Table 10.1 presents data from an experiment carried out by Wainwright et al. [18] to study nutritional effects on preweaning mouse pups. The level of nutrient availability was manipulated by rearing the pups in litter sizes ranging from three to twelve mice. On day 32, body weight and brain weight were measured. The values in the table are the average body weight (BODY) and brain weight (W) for each of 20 litters, two of each litter size (LITSIZ). For the purpose of illustration, we regard these averages as single observations. In this study, the effect of nutrition on brain weight is of particular interest.

Table 10.1. Average weight measurements from 20 litters of mice

Litter size	Body weight, g	Brain weight, g	Litter size	Body weight, g	Brain weight, g
3	9.447	0.444	8	7.040	0.414
3	9.780	0.436	8	7.253	0.409
4	9.155	0.417	9	6.600	0.387
4	9.613	0.429	9	7.260	0.433
5	8.850	0.425	10	6.305	0.410
5	9.610	0.434	10	6.655	0.405
6	8.298	0.404	11	7.183	0.435
6	8.543	0.439	11	6.133	0.407
7	7.400	0.409	12	5.450	0.368
7	8.335	0.429	12	6.050	0.401

Figure 10.2 consists of separate scatterplots of the average brain weight of the pups in a litter versus the corresponding average body weight and the litter size, as well as a third scatterplot of litter size versus average body weight. Notice that both body weight and litter size appear to be associated with the average brain weight for the mouse pups.

A simple linear regression model relates a response or outcome measurement such as brain weight to a single explanatory variable. Equation (10.1) is an example of a simple linear regression model. By comparison, a multiple linear regression model attempts to relate a measurement like brain weight to more than one explanatory variable. If we represent brain weight by W, then a multiple linear regression equation relating W to body weight and litter size would be

$$W = a + b_1(\mathrm{BODY}) + b_2(\mathrm{LITSIZ}).$$

At this point, it is helpful to introduce a little notation. If Y denotes a response variable and $X_1, X_2, \ldots, X_k$ denote explanatory variables, then a regression equation for Y in terms of $X_1, X_2, \ldots, X_k$ is

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

or

$$Y = a + \sum_{i=1}^{k} b_iX_i.$$

Remember, also, that if we wish to refer to specific values of the variables $X_1, X_2, \ldots, X_k$, it is customary to use the lower case letters $x_1, x_2, \ldots, x_k$.

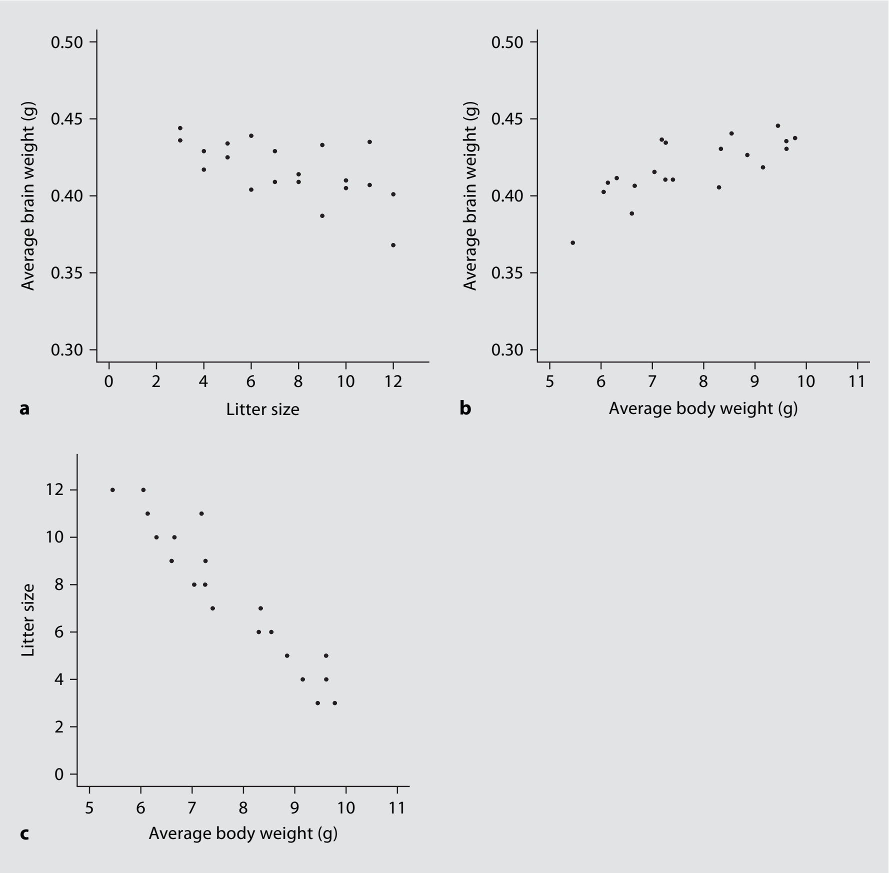


Fig. 10.2. Scatterplots of average weight measurements from 20 litters of mice. **a** Average brain weight vs. litter size. **b** Average brain weight vs. average body weight. **c** Litter size vs. average body weight.

A multiple linear regression analysis finds estimates $\hat{a}$, $\hat{b}_1, \ldots, \hat{b}_k$ of a, $b_1, \ldots, b_k$ which minimize the average value of

$$(Y - \hat{Y})^2 = (Y - \hat{a} - \sum_{i=1}^{k} \hat{b}_i X_i)^2.$$

The assumption which underlies this analysis is that, for specified covariate values $X_1 = x_1$, $X_2 = x_2$, …, $X_k = x_k$, the distribution of Y is normal with mean or expected value

$$a + \sum_{i=1}^{k} b_i \, x_i.$$

Table 10.2. Estimated regression coefficients and corresponding standard errors for simple linear regressions of brain weight on body weight and litter size

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
BODY	0.010	0.002	5.00	<0.0001
LITSIZ	–0.004	0.001	4.00	<0.001

The variances of the different normal distributions corresponding to different sets of covariate values are assumed to be the same. Notice that a regression analysis makes *no* assumptions about the distribution of the explanatory variables $X_1, X_2, \ldots, X_k$. In many designed experiments, values of the explanatory variables are, in fact, selected by the investigator in order to study the relationship between Y and $X_1, X_2, \ldots, X_k$ systematically.

In our example, two separate, simple linear regressions of brain weight on body weight and litter size can be performed. These lead to the two equations

$$\hat{W} = 0.336 + 0.010(\text{BODY}) \text{ and } \hat{W} = 0.447 - 0.004(\text{LITSIZ}).$$

If the multiplier, or regression coefficient, for any variable was zero, then that variable would have no influence on the response variable W. It is very unlikely that the best estimate of a regression coefficient would be exactly zero. A more reasonable question to ask is whether the data provide evidence to contradict the hypothesis that the regression coefficient could be zero, thereby confirming a relationship between the explanatory and response variables.

Table 10.2 lists the estimated regression coefficients and their corresponding standard errors for the two simple linear regressions of brain weight on body weight and litter size. If there is no relationship between the explanatory variable and brain weight, then the ratio $\hat{b}$/est. standard error ($\hat{b}$) has a Student's t distribution with the same number of degrees of freedom as the estimated standard error. The results of chapter 9, therefore, lead to the test statistic

$$T = \frac{|\hat{b}|}{\text{est. standard error } (\hat{b})}$$

to test the hypothesis that the regression coefficient b equals zero. To calculate the significance level of the test, the observed value of T is compared with the critical values of the modulus of a Student's t distribution with the appropriate degrees of freedom (cf. table 9.4). Table 10.2 also includes the observed values of T and the associated significance levels for each regression coefficient.

Table 10.3. Estimated regression coefficients and corresponding standard errors for a multiple linear regression of brain weight on body weight and litter size

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
BODY	0.024	0.007	3.43	0.003
LITSIZ	0.007	0.003	2.33	0.032

The significance levels summarized in table 10.2 indicate that the data provide strong statistical evidence of a relationship between brain weight and each of the explanatory variables. However, these relationships are not particularly strong physical effects. In fact, the average brain weight of a mouse pup increases by only 0.010 g per one-gram increase in body weight. Similarly, the average brain weight of a mouse pup decreases by 0.004 g for each unit increase in size of the litter into which the pup is born.

The analyses corresponding to the results presented in table 10.2 examine the separate effects of body weight and litter size on brain weight. The joint effects of the two variables are investigated via the multiple linear regression equation

$$W = a + b_1(\text{BODY}) + b_2(\text{LITSIZ}), \tag{10.3}$$

which is estimated by

$$\hat{W} = 0.178 + 0.024(\text{BODY}) + 0.007(\text{LITSIZ}).$$

Table 10.3 gives the estimated regression coefficients $\hat{b}_1$ and $\hat{b}_2$, their estimated standard errors and the ratios used to test for a non-zero coefficient, based on the model specified by equation (10.3).

Clearly, table 10.3 is quite different from table 10.2. The coefficients change in size and, for litter size, in sign. The difference arises because the test for a relationship between LITSIZ, for example, and brain weight in table 10.3 is performed when the other variable, BODY, is included in the model. Thus, although LITSIZ is inversely related to brain weight when examined singly (see table 10.2), the model results summarized in table 10.3 tell us that LITSIZ is positively related to brain weight after adjusting for the available information on body weight. On the other hand, body weight is positively related to brain weight, even after litter size is taken into account. Both covariates have coefficients which are significantly different from zero; therefore, each provides information concerning brain weight which is additional to that provided by the other variable. Thus, both covariates should be included in a model describing brain weight.

Table 10.3 is consistent with the biological concept of brain sparing, whereby the nutritional deprivation represented by litter size has a proportionately smaller effect on brain weight than on body weight. In a simple linear regression, litter size is negatively related to brain weight because the larger litters tend to consist of the smaller mice. In the multiple linear regression, the effect of body size is taken into account and the positive regression coefficient associated with litter size indicates that mice from large litters will have larger brain weights than mice of comparable size from smaller litters.

The analysis which we have described in this section is approximate and can be improved upon. The results of a linear regression analysis are frequently presented, in summary form, in an analysis of variance (ANOVA) table. Since this type of presentation does not extend easily to other regression models (see chapters 11–14), it is not of primary importance to the aims of this book. Our presentation is consistent with the usual analysis of other regression models which are widely used in medical research, and therefore serves as a useful introduction to the general topic of regression models.

Before we consider other types of regression models, we propose to briefly discuss correlation analysis, a historical antecedent to regression analysis. Also, the final section of this chapter describes ANOVA tables. This material is not needed to understand chapters 11–14, but is included for completeness. The amount of detail which is necessary to explain ANOVA tables far exceeds the complexity of most other explanations appearing in this book. We therefore suggest that the reader bypass §10.5 on a first reading.

10.4. Correlation

Regression models presuppose there is an outcome or response variable of some importance, and that interest in other variables derives from their potential influence on the outcome variable. Historically, the development of regression analysis was preceded by another approach known as correlation analysis.

Consider the case of two variables, Y and X. In a regression analysis, we assume Y has a normal distribution, but X may take any value and no distributional assumptions about X are required. Thus, X may be determined along with Y, X values may be fixed by the experimenter, e.g., X represents treatment received in a randomized clinical trial, or any number of factors might influence the X values observed in the data. Correlation analysis is restricted to the situation when X and Y are both random variables, and commonly assumes that both variables have a normal distribution.

Figure 10.2 shows separate plots of brain weight versus litter size and body weight, and litter size versus body weight. The plots indicate that when brain weight is high or low, body weight tends to be correspondingly high or low, whereas litter size tends to be the opposite. The relationship between litter size and body weight is similar to the relationship between brain weight and litter size. A numerical measure of the observed association between two variables is the correlation coefficient r. For completeness, we give the formula for r, which is

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where n is the number of paired observations on X and Y.

The correlation coefficient is a number between –1 and 1; the value r = 0 indicates there is no linear relationship between the two variables. A statistical procedure has been developed to test the hypothesis of no association between X and Y, based on the observed correlation coefficient, but we do not intend to discuss it here. If r is negative, then X and Y are said to be negatively correlated, implying that low values of X tend to occur with high values of Y and vice-versa. This kind of association is illustrated by the relationship between brain weight and litter size, which have a correlation coefficient of –0.62. If r is positive, then X and Y have a relationship like that observed between brain weight and body weight. The correlation coefficient for these variables is 0.75.

Figure 10.3 contains eight plots which illustrate the degree of linear relationship between the values of two variables, X and Y, corresponding to various positive and negative values of the correlation coefficient. The plots are based on 50 (X, Y) pairs and have been artificially constructed so that, in each case, the 50 X values displayed in each plot are identical.

The correlation coefficient can be used in the initial examination of a data set to identify relationships which deserve further study. However, it is generally more useful to think of linking two variables via a regression equation. From a regression analysis we can see very directly how changes in one variable are associated with changes in the other outcome variable. The regressions of litter size on brain weight and body weight are summarized by the equations

$$\text{LIT}\hat{\text{S}}\text{IZ} = 47.41 - 95.76(\text{W}) \text{ and } \text{LIT}\hat{\text{S}}\text{IZ} = 23.51 - 2.07(\text{BODY}).$$

Thus, for a specified value of brain weight or body weight, we could ‘predict’ a value for litter size.

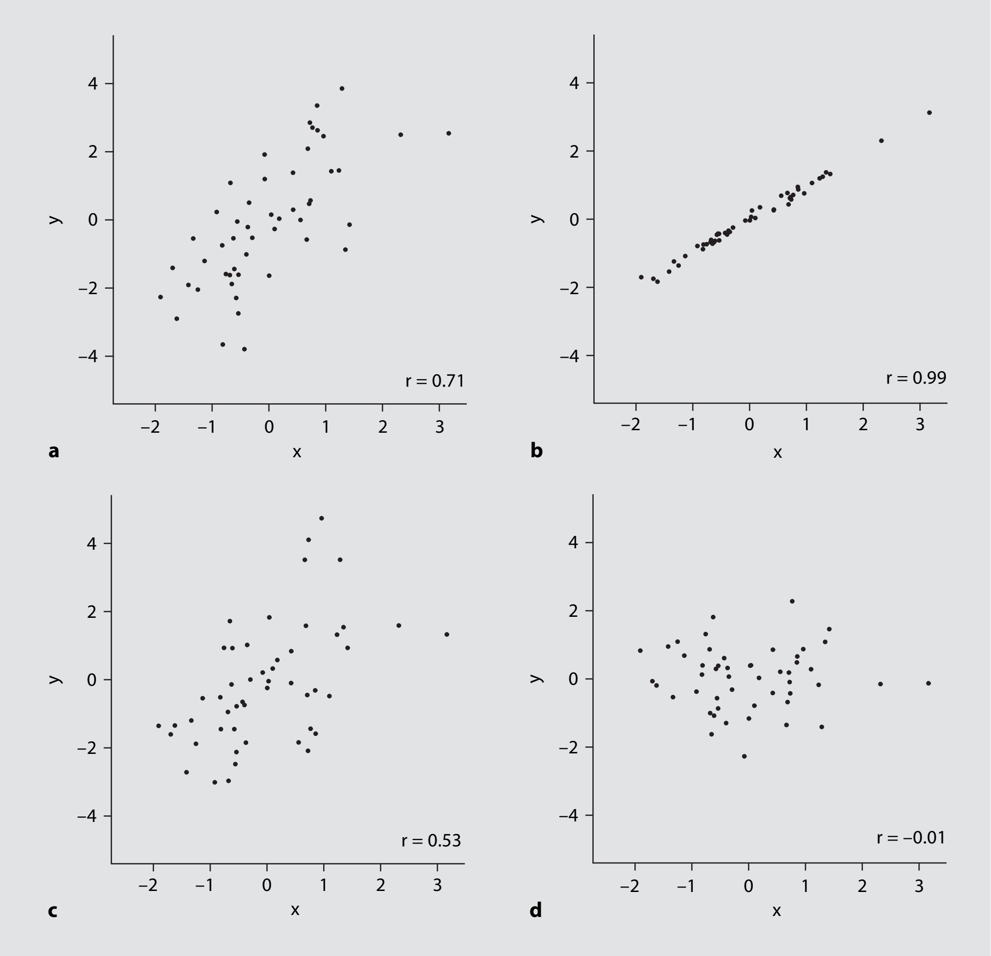
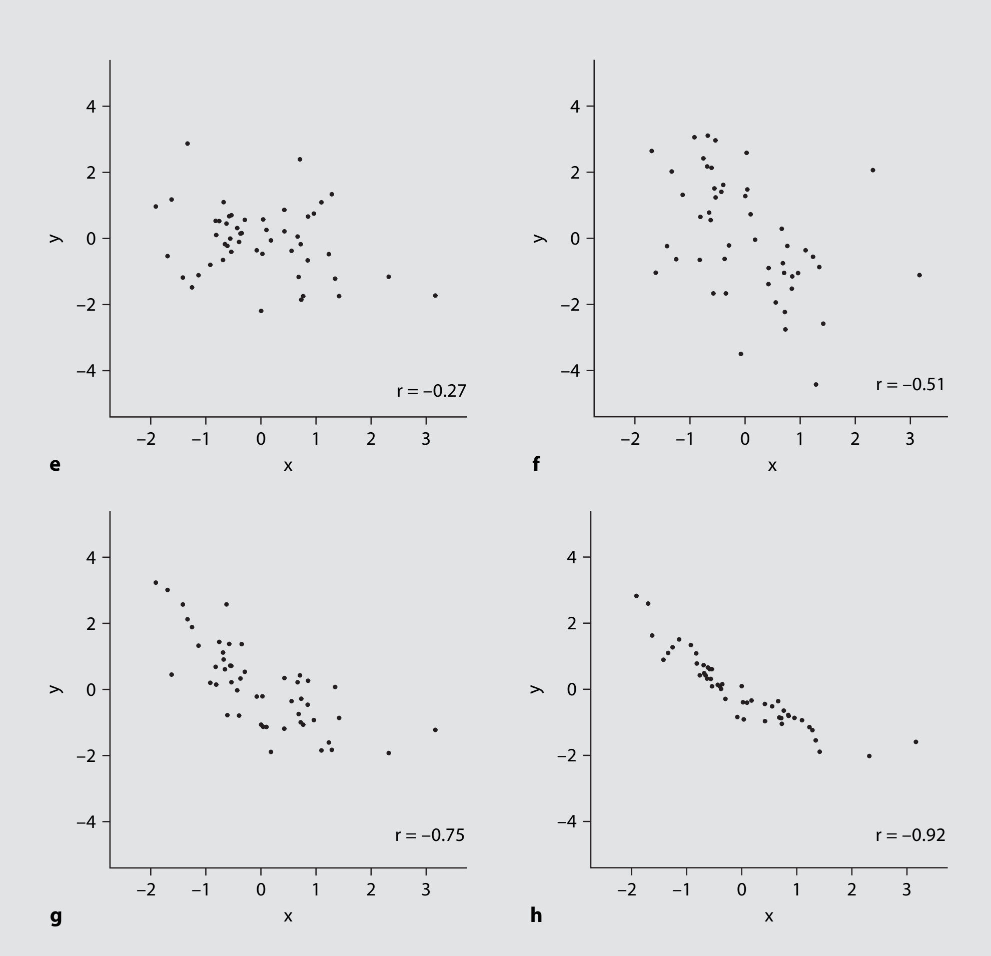


Fig. 10.3. Scatterplots of 50 artificial (X, Y) measurement pairs illustrating various values of the estimated correlation coefficient r. **a** r = 0.71, **b** r = 0.99, **c** r = 0.53, **d** r = −0.01, **e** r = −0.27, **f** r = −0.51, **g** r = −0.75, **h** r = −0.92.

Another reason for generally preferring regression analysis is the fact that correlation analysis is applicable only to situations when both X and Y are random. In many studies, the selection of subjects will depend on the values of certain variables for these subjects. Such selection invalidates correlation analysis. For example, Armitage et al. [19] indicate that if, from a large population of individuals, selection restricts the values of one variable to a limited range, the absolute value of the correlation coefficient will decrease.



Two additional points deserve brief mention. The first concerns correlation measures which do not depend on the assumption that X and Y have normal distributions. Two of these measures are Spearman's rank correlation coefficient and Kendall's τ (tau). These are useful for analyzing non-normal data, but the general reservations concerning correlation analysis which we have already mentioned apply equally to any measure of correlation.

The second issue is the following. Although we advocate the use of regression models, it is important to realize that the estimated regression line of Y on X is not the same as that for X on Y. This can be seen by comparing the regression lines for brain weight on litter size and litter size on brain weight. This asymmetry may seem strange, but it is linked to use of the differences $(Y - \hat{Y})$ – or $(X - \hat{X})$ if X is the response variable – as an estimation criterion. If one variable is random, say Y, and the other is selectively sampled, then the regression of Y on X will be sensible. Otherwise, the choice of a suitable regression model may be determined by some preference for predicting one variable on the basis of the other.

10.5. The Analysis of Variance

As we mentioned in §10.3, a multiple regression analysis is frequently summarized in an analysis of variance (ANOVA) table. This section, which is only necessary to the understanding of chapter 15, provides a brief introduction to ANOVA tables. On a first reading, we strongly advise readers to skip over this section and that related chapter and proceed to chapter 11.

In the example we have been discussing, we have observed values of the response variable, W, and predicted values, $\hat{W}$, which are derived from the regression equation. Let $\overline{W}$ represent the average of all the observed brain weights. In chapter 1, we learned that measures of spread, such as variance, usually involve the quantity $(W - \overline{W})$, which represents the difference of observed values of W from their average. Strictly speaking, we should refer to $(W - \overline{W})^2$ rather than the simple difference $(W - \overline{W})$. We will shortly introduce the squared difference; however, it is simpler, for the moment, to begin our explanation by discussing $(W - \overline{W})$. In fact, the analysis of variance is based on the relation

$$(W - \overline{W}) = (W - \hat{W}) + (\hat{W} - \overline{W}),$$

which decomposes $(W - \overline{W})$ into two separate components. If we regard $(W - \overline{W})$ as 'total variation', i.e., variance, it is natural to ask what components of variation $(W - \hat{W})$ and $(\hat{W} - \overline{W})$ represent.

In the absence of a regression model for the mean value of W, the dependent variable, $\overline{W}$ represents a natural estimate. Thus, $(W - \overline{W})$ can be interpreted as the variation of W around this natural estimate. Once we have specified a regression model for the mean value of W, we have a different estimate, $\hat{W}$, which is based on the model. Moreover, $(W - \hat{W})$ will generally be smaller than $(W - \overline{W})$. Therefore, of the total variation represented by $(W - \overline{W})$, the component $(\hat{W} - \overline{W})$ has been 'explained' or accounted for by the fitted regression model for W. That is, since $\hat{W}$ is now the estimated mean value of W, based on the values of LITSIZ and BODY, we expect to see W differ from $\overline{W}$ by $(\hat{W} - \overline{W})$. The remaining component, $(W - \hat{W})$, is called the 'residual variation', i.e., the variation in W which is not explained by the regression equation.

Now that we have informally described the decomposition of the variance in W, we introduce squared differences, which are the correct representation. Let W_1, W_2, ..., W_n be n observed values of W, and let $\overline{W} = \frac{1}{n} \sum_{i=1}^{n} W_i$ be their average. The total variation in these data can be represented by

$$\sum_{i=1}^{n}(W_i - \overline{W})^2 = (W_1 - \overline{W})^2 + (W_2 - \overline{W})^2 + \ldots + (W_n - \overline{W})^2,$$

which is (n–1) times the sample variance for W. Although it is not obvious, it can be shown that

$$\sum_{i=1}^{n}(W_i - \overline{W})^2 = \sum_{i=1}^{n}(W_i - \hat{W}_i)^2 + \sum_{i=1}^{n}(\hat{W}_i - \overline{W})^2.$$

Thus, the total variation in W decomposes into two sums of squared differences. In view of the preceding discussion involving simple differences, it should not be difficult to accept that the second sum represents the component of total variation in W which is accounted for by the fitted regression model, while the first sum represents the residual variation. If we use SS to represent a sum of squares, then we can rewrite the above equation, symbolically, as

$$SS_{Total} = SS_{Residual} + SS_{Model};$$

the reasons for the subscripts should be quite obvious.

Each sum of squares of the form

$$\sum_{i=1}^{n}(W_i - [\text{estimated value of } W_i])^2$$

has associated with it a quantity called its degrees of freedom (DF). This quantity is equal to the number of terms in the sum minus the number of values which must be calculated from the W_i's in order to specify all the estimated values. For example, in SS_{Total} the estimated value for each W_i is the same, namely $\overline{W}$. Thus, we need one calculated value, and the degrees of freedom for SS_{Total} are (n – 1). For $SS_{Residual}$, the estimated value of each W_i is equal to

$$\hat{W}_i = \hat{a} + \sum_{j=1}^{k} \hat{b}_j x_j,$$

where x_1, …, x_k are the values of the covariates corresponding to W_i. These n estimates require (k + 1) calculated values, $\hat{a}$, $\hat{b}_1$, ..., $\hat{b}_k$; therefore, the degrees of freedom for $SS_{Residual}$ are (n – k – 1).

Although we shall not attempt to justify the following result, it can be shown that

$$DF_{Total} = DF_{Residual} + DF_{Model}.$$

Table 10.4. An ANOVA table for the regression analysis of brain weight

Term	SS	DF	MS	F
Model	0.004521	2	0.002260	15.8
Residual	0.002429	17	0.000143	
Total	0.006950	19		

$R^2 = 0.65$.

Therefore, since DF_{Total} equals (n – 1) and $DF_{Residual}$ equals (n – k – 1), it follows that the degrees of freedom for SS_{Model} are equal to k.

The only remaining unknown quantities which appear in an ANOVA table are called mean squares (MS); these are defined to be the ratio of a sum of squares to its degrees of freedom, i.e., MS = SS/DF. The variance estimates which we discussed in chapter 9 can be recognized as mean squares, and this is partly the reason that statisticians call the method of analysis which we are presently describing 'the analysis of variance'.

The typical format for an ANOVA table is shown in table 10.4, which gives the appropriate entries for the regression of W on BODY and LITSIZ. From this table, two quantities are usually calculated. The first of these is a ratio called R^2 (R-squared), which is equal to

$$R^2 = SS_{Model}/SS_{Total}.$$

This ratio indicates the fraction of the total variation in W which is accounted for by the fitted regression model. There are no formal statistical tests associated with R^2, and it is primarily used for information purposes only. In table 10.4, the value of R^2 is 0.65, indicating that 65% of the total variation in W is explained by the regression model.

The second quantity which is calculated from an ANOVA table such as table 10.4 is the ratio of MS_{Model} to $MS_{Residual}$. This number is frequently called an F-ratio because, if the null hypothesis that *all* the regression coefficients b_1, …, b_k are equal to zero is true, the ratio $MS_{Model}/MS_{Residual}$ should have an F-distribution with DF_{Model} and $DF_{Residual}$ degrees of freedom. Recall that the F-distribution was introduced in §9.5.

The observed value of this F-ratio is usually compared with the critical values for the appropriate F-distribution (see table 9.7), and if the observed value exceeds the critical value, the regression is said to be significant, i.e., the data contradict the null hypothesis that all the regression coefficients are zero. In table 10.4, the observed F-ratio is 0.00226/0.000143 = 15.8. Since the 5%

Table 10.5. An expanded ANOVA table for the regression analysis of brain weight

Term	SS	DF	MS	F
Model	0.004521	2	0.002260	15.8
BODY	0.003869	1	0.003869	27.1
LITSIZ|BODY	0.000652	1	0.000652	4.6
Residual	0.002429	17	0.000143	
Total	0.006950	19		

critical value for an F-distribution with 2 and 17 degrees of freedom is 3.59, we conclude that the regression is significant at the 5% level; the data contradict the hypothesis that both BODY and LITSIZ have no effect on brain weight. This result is consistent with the conclusion which we reached in §10.3 (cf. table 10.3) that the regression coefficients for BODY and LITSIZ are significantly different from zero.

Table 10.3 summarizes the results of tests concerning the effect of individual covariates when the other covariate was included in the regression model. For example, we determined whether LITSIZ had any influence on brain weight if BODY was already included in the model. By comparison, the F-test which we have discussed above is a joint test of the hypothesis that all the regression coefficients are zero. However, the ANOVA table which we have described can be generalized to address the question of the individual effect of each covariate.

If we had calculated an ANOVA table for the regression of W on BODY, then the SS_{Model} would have been 386.9×10^{-5}, and DF_{Model} would have been one. This number, 386.9×10^{-5}, represents the component of the SS_{Model} in table 10.4 which is due to BODY alone. If we next add LITSIZ to the regression equation, so that we are fitting the model described by table 10.4, the SS_{Model} increases by 65.2×10^{-5}, from 386.9×10^{-5} to 452.1×10^{-5}. Thus, 65.2×10^{-5} is the component of SS_{Model} which is accounted for by LITSIZ, when the model already includes BODY. Notice, also, that we could have performed these calculations in the reverse order, i.e., LITSIZ first, followed by BODY.

In table 10.5, the SS_{Model} is divided into two parts. One part is labelled BODY, and represents the component of SS_{Model} which is due to BODY; the other part is labelled LITSIZ|BODY, and represents the component which is due to LITSIZ in addition to BODY. The degrees of freedom for each component in the sum of squares are one since DF_{Model} equals two and the component of the SS_{Model} which is due to BODY has one degree of freedom. Therefore, two F-ratios can be calculated, and these appear in the last column of table 10.5;

Table 10.6. An ANOVA table for the regression of brain weight on litter size

Term	SS	DF	MS	F
Model	0.002684	1	0.002684	11.3
Residual	0.004266	18	0.000237	
Lack of fit	0.001290	8	0.000161	0.54
Pure error	0.002976	10	0.000298	
Total	0.006950	19		

one F-ratio corresponds to BODY alone, and the other represents LITSIZ|BODY. If the null hypothesis that BODY has no influence on W is true, the F-ratio for BODY should have an F-distribution on 1 and 17 degrees of freedom. Quite separately, if the null hypothesis is true that LITSIZ has no influence on W which is additional to the effect of BODY, then the F-ratio for LITSIZ|BODY should also have an F-distribution on 1 and 17 degrees of freedom. Since the 5% critical value for this particular F-distribution is 4.45, we conclude that each of the terms in the regression model – BODY and LITSIZ|BODY – is necessary since a test of the respective null hypothesis has a significance level of less than 5%. This conclusion coincides with the analysis which we discussed in §10.3 (cf. table 10.3).

To illustrate one additional type of calculation, table 10.6 presents an ANOVA table for the simple linear regression of brain weight on litter size. We assume no additional variables are available. In an analysis of variance, the $MS_{Residual}$ is often used as an estimate of variance. The $MS_{Residual}$ represents the variation which cannot be accounted for by the regression equation, and it is often assumed that this unexplained variation is the natural variance of the observations about the estimated values which are determined by the regression equation. Since we have two independent observations on each of the ten litter sizes, we can calculate a separate, independent estimate of the natural variation in the model. Let W_1 and W_2 represent the two observations for a single litter size. The natural estimate of the mean brain weight for this litter size is $\overline{W} = (W_1 + W_2)/2$, and an estimate of the residual variation, based on these weights alone, would be a SS which is equal to $(W_1 - \overline{W})^2 + (W_2 - \overline{W})^2$. Since we have two terms in the sum and one estimate, $\overline{W}$, the degrees of freedom for this sum would be one. If we repeat this calculation for all ten litter sizes, then the total of the ten individual sums of squares is 297.6×10^{-5}, and this total sum of squares would have $10 \times 1 = 10$ degrees of freedom.

This type of calculation, which leads to a MS of $297.6 \times 10^{-5}/10 = 29.8 \times 10^{-5}$ in our example, is often called a calculation of 'pure error'. This is because, regardless of the information we have concerning an individual litter size, the variation of independent observations from litters of the same size can never be accounted for by the regression equation. Therefore, $MS_{\text{Pure Error}}$ is a better estimate of the natural variance of the observations than MS_{Residual}. Recall that we are assuming we do not have any information apart from brain weight and litter size. To calculate a $SS_{\text{Pure Error}}$ for the regression summarized in table 10.5 would require independent observations on mice which have the same body weight and were reared in litters of the same size. Such observations are not available, and therefore the pure error calculations cannot be carried out in that particular case.

The $SS_{\text{pure Error}}$ is one component of SS_{Residual}, and the remainder is usually called the Lack of Fit component, since it constitutes a part of the SS_{Total} which cannot be accounted for, either by the regression model or by pure error. The $SS_{\text{Lack of Fit}}$ measures the potential for improvement in the model for W if additional covariate information is used, or possibly if an alternative form of the regression equation is specified instead of the current version. A test that the lack of fit in the current model is significant can be based on the F-ratio $MS_{\text{Lack of Fit}}/MS_{\text{pure Error}}$. If the null hypothesis that there is no lack of fit in the current model is true, this ratio should have an F-distribution with degrees of freedom $DF_{\text{Lack of Fit}} = DF_{\text{Residual}} - DF_{\text{Pure Error}}$ and $DF_{\text{Pure Error}}$. We will not discuss the question of lack of fit in detail; however, if there is a significant lack of fit, it is customary to plot the values of $W - \hat{W}$ for all the observations, to see if they appear to be normally distributed. If so, then the lack of fit is usually attributed to unavailable information rather than the existence of better alternative models which involve the same variates. In table 10.6, the test for Lack of Fit is not significant.

Despite its very obvious link to the methods of linear regression, ANOVA is frequently regarded by many investigators as a specialized statistical method. This is particularly the case when ANOVA is used to evaluate the results of carefully designed experiments in which the researcher fixes or controls the operational settings of certain explanatory variables, which are usually called factors. Although ANOVA, and the corresponding ANOVA tables that we have described in this section, do not play a role in the methods of analysis used with other types of regression models such as those involving response measurements that are binary or count data, the use of ANOVA is occurring with greater frequency in the medical literature. Therefore, we will return to the topic of ANOVA in chapter 15, but judge it best to introduce first the use of regression methods for other kinds of response measurements in chapters 11–14.

11

Binary Logistic Regression

11.1. Introduction

One reason that linear regression is not as widely used in medical statistics as in other fields is that the outcome variable in a medical study frequently cannot be assumed to have a normal distribution. A common type of response or outcome variable in medical research is a binary variable. These response variables take one of two values, and were discussed extensively in chapters 2 through 5. In this chapter, we describe a particular regression model for a binary response variable.

To illustrate the methodology, we shall consider an example concerning bone marrow transplantation for the treatment of aplastic anemia. One response of major interest is graft rejection. A binary variable, Y, can be defined so that Y = 1 corresponds to graft rejection and Y = 0 represents graft acceptance.

Table 11.1 is taken from an article by Storb et al. [3]. It presents a binary logistic regression analysis of graft rejection, relating the response variable, Y,

Table 11.1. Maximum likelihood fit of a binary logistic regression model to marrow-graft rejection data on 68 patients with aplastic anemia [3]

Factor	Logistic coefficient	Standard error	Ratio
Marrow cell dose	–1.005	0.344	–2.92 ($p < 0.01$)
Age	–0.457	0.275	–1.66 ($p < 0.10$)
Blood units	1.112	0.672	1.65 ($p < 0.10$)
Transplant year	0.735	0.319	2.30 ($p < 0.05$)
Androgen treatment	1.417	0.805	1.76 ($p < 0.10$)

Reprinted from Storb et al. [3] with the kind permission of the publisher.

to marrow cell dose in units of 10^8 cells per kilogram of body weight, patient age in decades, extent of prior blood units transfused, transplant year (minus 1970) and preceding androgen treatment. The prior blood unit variable was zero if the patient received less than 10 whole blood units prior to transplantation, and one otherwise. The androgen variable was zero if the patient had not received androgen previously, and one otherwise. Let us call these five variables X_1, X_2, X_3, X_4, and X_5. A convenient, shorthand notation for the set of variables $\{X_1, X_2, X_3, X_4, X_5\}$ is $\underset{\sim}{X}$. Remember, also, that we denote a particular value for a variable by the corresponding lower case letter, viz $\underset{\sim}{x} = \{x_1, x_2, x_3, x_4, x_5\}$.

11.2. Logistic Regression

Since Y can assume only two possible values, it would be unrealistic to entertain a linear regression model such as

$$Y = a + b_1X_1 + \ldots + b_5X_5 = a + \sum_{i=1}^{5} b_iX_i.$$

Theoretically, the right-hand side of this equation can take any value between minus infinity ($-\infty$) and plus infinity ($+\infty$) unless we restrict the values of a and the regression coefficients $b_1, \ldots, b_5$.

In a linear regression model, the expression $a + \Sigma b_iX_i$ is assumed to be the expected value of a normal distribution. The expected value of a binary variable such as Y turns out to be the probability that Y = 1. Thus, it is more reasonable to consider a regression model which involves the probability of graft rejection, i.e., Pr(Y = 1). A probability lies between zero and one, and this is still too narrow a range of values for the expression $a + \Sigma b_iX_i$. However, if a probability, say p, is between zero and one, then $p/(1 - p)$ belongs to the interval $(0, \infty)$ and $\log\{p/(1 - p)\}$ belongs to the interval $(-\infty, \infty)$. This is the same range of values to which the expression $a + \Sigma b_iX_i$ belongs.

If we represent the probability of graft rejection, Y = 1, by $\Pr(Y = 1|\underset{\sim}{x})$ for an individual with covariate values $\underset{\sim}{x}$, then a binary logistic regression model for Y is specified by the equation

$$\log\left\{\frac{\Pr(Y=1|\underset{\sim}{x})}{1-\Pr(Y=1|\underset{\sim}{x})}\right\} = \log\left\{\frac{\Pr(Y=1|\underset{\sim}{x})}{\Pr(Y=0|\underset{\sim}{x})}\right\} = a + \sum_{i=1}^{5} b_ix_i.$$

An equivalent way of specifying the model is via the equation

$$\Pr(Y=1|\underset{\sim}{x}) = \frac{\exp(a + \sum_{i=1}^{5} b_ix_i)}{1 + \exp(a + \sum_{i=1}^{5} b_ix_i)},$$

Table 11.2. Graft rejection status and marrow cell dose data for 68 aplastic anemia patients

Graft rejection	Marrow cell dose (10^8 cells/kg)		Total
	<3.0	≥3.0	
Yes	17	4	21
No	19	28	47
Total	36	32	68

which reveals that the model links the linear expression $a + \Sigma b_i x_i$ to the probability of graft rejection. Now, as in linear regression, if b_i is zero, then the factor represented by X_i is not associated with graft rejection. As in the case of linear regression analysis (see chapter 10), a suitable statistic for testing the hypothesis that the regression coefficient, b_i, equals zero is

$$T = \frac{|\hat{b}_i|}{\text{est. standard error}(\hat{b}_i)}.$$

Occasionally, the results of an analysis may be presented in terms of the ratio

$$\frac{\hat{b}_i}{\text{est. standard error}(\hat{b}_i)},$$

which is equal to T, apart from the sign. This is the situation in table 11.1. In either case, the conclusion regarding the covariate represented by X_i is the same.

In table 11.1, the largest ratio is associated with marrow cell dose ($p = 0.004$). Transplant year has a significant effect on graft rejection, with a p-value of 0.02. The other covariates are not significant at the 5% level, although the associated p-values are all less than 0.10. Remember that a test of the hypothesis that a certain regression coefficient is zero is a test for the importance of the corresponding covariate, having adjusted for all the other variables in the regression model. For example, the effect of transplant year cannot be attributed to a change in marrow cell dose values with time, since marrow cell dose is included in the model when we test the hypothesis $b_4 = 0$, i.e., the covariate representing transplant year is not associated with graft rejection.

Details of the calculations that are involved in estimating a binary logistic regression model, and that are known as maximum likelihood estimation, are beyond the scope of this book. To actually use this methodology to analyze a particular set of data, it would be necessary to consult a statistician. However, we hope our brief discussion of the binary logistic regression model has been

Table 11.3. A logistic regression analysis of graft rejection and marrow cell dose in 68 aplastic anemia patients

Regression coefficient	Estimate	Estimated standard error	Test statistic
a	–1.95	0.53	–
b	1.83	0.63	2.90 (p = 0.004)

informative, and will permit readers to appraise the use of this technique in published papers critically.

There are many aspects which are common to the use of quite different regression models. For this reason, we have minimized our discussion of binary logistic regression; much of the discussion in chapters 12–15 is equally relevant to logistic regression. Nevertheless, as another illustration of this methodology, and as a means of addressing a topic which we have thus far neglected, we discuss the application of logistic regression to 2 × 2 tables in the next section.

11.3. Estimation in 2 × 2 Tables

The discussion of 2 × 2 tables in chapters 2 through 5 concentrates on the concept of a significance test. This emphasis was adopted for pedagogical purposes, and we now turn to the equally important problem of estimation in 2 × 2 tables. This can be done within the framework of the binary logistic regression model.

Consider the data presented in table 11.2 concerning graft rejection in 68 aplastic anemia patients; each marrow cell dose is recorded as being one of two types, namely either less than or at least 3.0×10^8 cells/kg. Let Y represent graft rejection as in §§11.1, 11.2 and let X be a binary covariate, where X = 1 corresponds to a low marrow cell dose and X = 0 indicates a higher dose. A binary logistic regression model for graft rejection and marrow cell dose is specified by the equation

$$\Pr(Y=1\,|\,x)=\frac{\exp(a+bx)}{1+\exp(a+bx)}. \tag{11.1}$$

Therefore, the probability of graft rejection for a high marrow cell dose (X = 0) is exp(a)/{1 + exp(a)}; the corresponding probability for the lower dose (X = 1) is exp(a + b)/{1 + exp(a + b)}.

The estimation of b is of major importance in studying the influence of marrow cell dose on graft rejection. Table 11.3 presents the estimation of model (11.1) for the data in table 11.2. A test of the hypothesis that b, the regression coefficient, equals zero is based on the observed value of the test statistic T = $|\hat{b}|/\{\text{est. standard error}(\hat{b})\}$ which equals 2.90. Since this observed value exceeds the 5% critical point given in table 8.1, there is evidence to contradict the hypothesis that marrow cell dose does not influence graft rejection, i.e., the hypothesis that b equals zero.

The larger $\hat{b}$ is, the larger is the estimated effect of a low marrow cell dose on graft rejection. As we saw in chapter 8, $\hat{b}$ is only a single number, and if we wish to estimate b, a confidence interval should also be calculated. A 95% confidence interval for b is defined to be

$$\hat{b} \pm 1.96\{\text{est. standard error}(\hat{b})\},$$

and can be represented by the interval (b_L, b_H).

The simplest way to think about b is in terms of an odds ratio. If p represents the probability of graft rejection, then p/(1 – p) is called the odds in favor of rejection. Now, let p_1 represent the probability of rejection for the higher marrow cell dose and p_2 the corresponding probability for the lower dose; then the ratio

$$\frac{p_2/(1-p_2)}{p_1/(1-p_1)}$$

is called the odds ratio (OR). If equation (11.1) is used to define p_1 and p_2, then it turns out that

$$\text{OR} = e^{b}.$$

Since a 95% confidence interval for b is (b_L, b_H), the corresponding interval for the odds ratio, OR, is

$$(e^{b_L}, e^{b_H}),$$

with an estimate of OR being $\widehat{\text{OR}} = \exp(\hat{b})$.

For a simple 2 × 2 table, the formula for $\hat{b}$ can be stated explicitly. If the table is of the form shown in table 11.4, then

$$\hat{b} = \log\left(\frac{ps}{qr}\right) \text{ and } \widehat{\text{OR}} = \frac{ps}{qr}.$$

In addition, an approximation to the estimated standard error of $\hat{b}$, based on the logistic regression model, is

$$\left(\frac{1}{p}+\frac{1}{q}+\frac{1}{r}+\frac{1}{s}\right)^{1/2}.$$

Table 11.4. The format of a 2 × 2 table in which the estimate of the odds ratio is ps/qr; the symbols – and + indicate absence and presence, respectively

Factor 2	Factor 1	
	–	+
–	p	q
+	r	s

This estimate can be somewhat too small, but it is convenient for quick calculation. In our example, the approximate value is 0.63, which is equal to the estimate of 0.63 from the logistic regression analysis.

In chapter 5, we discussed the use of stratification to adjust the test for no association in a 2 × 2 table for possible heterogeneity in the study population. In general, regression models are designed to make such adjustments more efficiently, but the stratification approach can be viewed as a special case of a regression model.

In the regression model for graft rejection, represented by the equation

$$\Pr(Y=1\,|\,x)=\frac{\exp(a+bx)}{1+\exp(a+bx)},$$

the parameter a determines the probability of graft rejection in individuals with X = 0, while b measures the change in this probability if X = 1. Other factors relevant to graft rejection would alter the overall probability of rejection in subgroups of the data. For example, table 11.2 can be subdivided into two 2 × 2 tables by stratifying according to transplant year (After 1972 – No or Yes). This is done in table 11.5. If these two tables are numbered 1 and 2, then we would define the two logistic regression models

$$\Pr(Y=1\,|\,x)=\begin{cases}\dfrac{\exp(a_1+bx)}{1+\exp(a_1+bx)}\ \text{for table 1,}\\[2ex]\dfrac{\exp(a_2+bx)}{1+\exp(a_2+bx)}\ \text{for table 2.}\end{cases}\tag{11.2}$$

In these models, the probability of rejection changes from table 1 to table 2 because the parameter a varies; however, the odds ratio parameter b, which measures the association between marrow cell dose and graft rejection, is assumed not to change. This type of model underlies the approach to combining 2 × 2 tables which we described in chapter 5. For any subclassifications of the popu-

Table 11.5. Graft rejection and marrow cell dose data in 68 aplastic anemia patients stratified by year of transplant

Graft rejection	Transplant year after 1972					
	no marrow cell dose (10^8 cells/kg)			yes marrow cell dose (10^8 cells/kg)		
	<3.0	≥3.0	total	<3.0	≥3.0	total
Yes	4	2	6	13	2	15
No	9	16	25	10	12	22
Total	13	18	31	23	14	37
	Table 1			Table 2		

Table 11.6. A logistic regression analysis of graft rejection and marrow cell dose, stratified by year of transplant

Regression coefficient	Estimate	Estimated standard error	Test statistic
a_1	–2.37	0.65	–
a_2	–1.54	0.59	–
b	1.72	0.64	2.69 (p = 0.007)

lation, including matched pairs, we can specify logistic regression models for each subgroup by using different a parameters and the same b parameter. Based on these models, b is estimated in order to study the association of interest.

Table 11.6 presents the estimation of model (11.2). The estimate of b is 1.72 and the observed value of the statistic used to test for no association is 1.72/0.64 = 2.69. This result is consistent with the unstratified analysis which we discussed earlier.

Table 11.6 also records estimates of a_1 and a_2. As the number of subgroups becomes large, problems do arise in estimating the a parameters. Specialized methodology for estimating b, alone, does exist and should be used in these situations; however, the details of this methodology are beyond the scope of this book. For our purposes, the nature of the logistic regression model, and the use of $\hat{b}$ to test for association, are more important.

Finally, we note that it is possible to test the assumption that the odds ratio, exp(b), is the same in the stratified 2 × 2 tables. If there are only a few tables,

Table 11.7. Data from a study of fetal mortality and prenatal care (L = Less, M = More) in two clinics

Prenatal care	Clinic 1		Clinic 2	
	L	M	L	M
Died	12	16	34	4
Survived	176	293	197	23

we can calculate separate interval estimates of b for each table and see if they overlap. More formal tests are rather complicated, but should be carried out if the assumption that b is constant is thought to be questionable in the least. A statistician should be consulted concerning these tests, and the appropriate way to proceed with the analysis, if the assumption that b is the same in the stratified 2 × 2 tables is not supported by the data.

Comment:

Readers who dip into the epidemiological literature will observe that logistic regression is frequently being used to analyze case-control studies. In this literature, it is common to see exp(b) referred to as a relative risk. For the model defined in equation (11.1), the relative risk associated with a low marrow cell dose would be $\Pr(Y = 1|X = 1)/\Pr(Y = 1|X = 0)$, or p_2/p_1 in our later notation. For a rare disease, the type usually investigated via a case-control study, p_1 and p_2 are small and therefore $1 - p_1$ and $1 - p_2$ are close to 1. In this situation, there is little difference between the odds ratio $\{p_2/(1 - p_2)\}/\{p_1/(1 - p_1)\}$ and the relative risk p_2/p_1. Therefore, epidemiologists frequently ignore the approximation which is involved, and refer to estimates of odds ratios from a case-control study as estimates of relative risks. It is also worth noting that the application of logistic regression to case-control studies involves certain arguments which go beyond the scope of this book. However, the nature of the conclusions arising from such an approach will be as we have presented them in this chapter.

11.4. Reanalysis of a Previous Example

In chapter 5, we discussed data concerning fetal mortality and prenatal care (L ≡ less, M ≡ more) in two clinics. The data are summarized in the 2 × 2 tables shown in table 11.7. Logistic regression analyses of these data, based on the unstratified model (see equation 11.1) and the stratified model

Table 11.8. Two logistic regression analyses of fetal mortality and prenatal care

	Regression coefficient	Estimate	Estimated standard error	Test statistic
Unstratified model	a	–2.76	0.23	–
	b	0.67	0.28	2.39 (p = 0.017)
Stratified model	a_1	–2.88	0.24	–
	a_2	–1.89	0.35	–
	b	0.15	0.33	0.45 (p = 0.65)

(see equation 11.2), are given in table 11.8. As we found in chapter 5, the unstratified model indicates there is a significant association (p = 0.017), whereas the more appropriate stratified analysis suggests that there is no association (p = 0.65) between fetal mortality and the amount of prenatal care received.

Tables 11.3, 11.6 and 11.8 illustrate that a stratified analysis, although generally appropriate, may or may not lead to different conclusions than an unstratified analysis. It is the potential for different conclusions that makes the adjustment for heterogeneity in a population important.

11.5. The Analysis of Dose-Response Data

As we have already seen in previous sections of this chapter, the binary logistic regression model is ideal for analyzing the dependence of a binary response variable on a set of explanatory variables, or covariates. Therefore, we wish to emphasize that the method of analysis which is discussed in this section is simply a special case of binary logistic regression. However, it represents a situation that is common in clinical studies. Moreover, the clinical example which we intend to discuss involves certain aspects of regression models which have not arisen in any of the examples we have previously considered.

Duncan et al. [20] report the results of a study which was initiated to investigate the effect of premedication on the dose requirement in children of the anaesthetic thiopentone. The study involved observations on 490 children aged 1–12 years. These patients were divided into four groups, three of which received different types of premedication. No premedication of any kind was administered to the fourth group of patients. All the children subsequently received an injection of 2.0–8.5 mg/kg of thiopentone in steps of 0.5 mg/kg. The anaesthetic was administered to each patient over a 10-second interval, and the eyelash reflex was tested 20 seconds after the end of the thiopentone

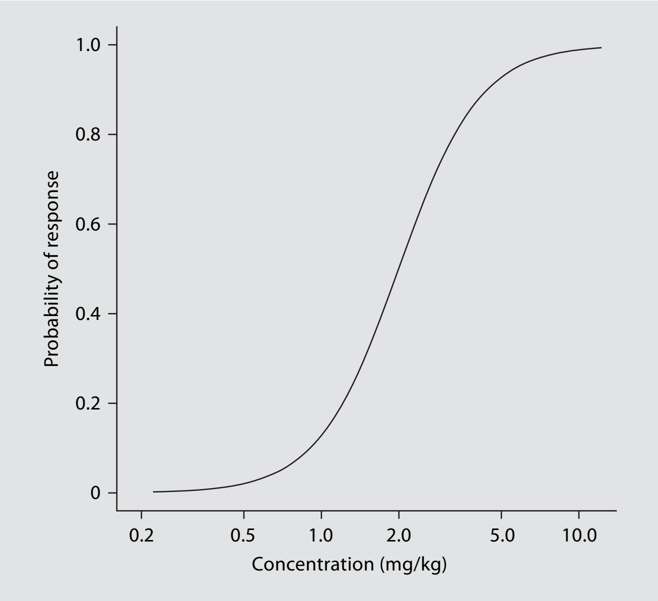


Fig. 11.1. A graph showing how the probability of responding might depend on the logarithm of the concentration in a dose-response study.

injection. If the eyelash reflex was abolished, the patient was deemed to have responded to the anaesthetic.

The investigation described above is typical of a class of clinical studies involving a binary response variable. Clearly, the purpose of the research is to assess the dependence of the response on a continuous variable which is under the control of the researcher. Investigations of this type are generally referred to as dose-response studies, because the clinician administers a measured concentration of a particular substance to each subject in a sample and then observes whether or not the subject exhibits the designated response. A principal assumption on which dose-response studies are based is the notion that the probability of responding depends in a simple, smooth way on the concentration. Figure 11.1 shows an example of this smooth relationship. In addition to estimating this dependence, researchers are usually interested in questions which concern differences in the dependence on concentration among well-defined subgroups of the population.

The method of binary logistic regression, which we introduced in §§11.1 and 11.2, is ideally suited to the analysis of dose-response data. In the use of this regression model to analyze data from such a study, it is common practice to choose the logarithm of the measured concentration as the explanatory

Table 11.9. Maximum likelihood fit of a binary logistic regression model to data on 137 children premedicated with TDP and atropine and then anaesthetized with thiopentone

Regression coefficient	Estimate	Estimated standard error	Test statistic
a	–1.92	0.82	–
b	2.78	0.72	3.86 (p = 0.0001)

variable rather than the actual concentration. Since use of the logarithmic value tends to improve the fit of the model to the data, we will follow this convention and represent the log concentration, or dose, of the administered substance by the letter d.

Let the binary variable Y denote the observed response, with Y = 1 indicating occurrence of the event of interest and Y = 0 its absence. Then Pr(Y = 1|d) represents the probability of observing the designated response in a subject who receives dose d. As we saw in §11.2, a binary logistic regression model for Y is specified by the equation

$$\log\left\{\frac{\Pr(Y=1\,|\,d)}{1-\Pr(Y=1\,|\,d)}\right\} = \log\left\{\frac{\Pr(Y=1\,|\,d)}{\Pr(Y=0\,|\,d)}\right\} = a + bd,$$

which is equivalent to requiring that

$$\Pr(Y=1\,|\,d) = \frac{e^{a+bd}}{1+e^{a+bd}}.$$

If b, the regression coefficient for dose, is zero, then dose, and hence concentration, is not associated with the probability of a response. In §11.2 we indicated that a suitable statistic for testing the hypothesis that b equals zero is

$$T = \frac{|\hat{b}|}{\text{est. standard error}(\hat{b})}.$$

Table 11.9 presents an analysis of a subset of the dose-response data which were collected for the study described by Duncan et al. [20]. The data were made available by Mr. B. Newman. The results summarized in the table pertain solely to the group of 137 patients who were premedicated orally with TDP (trimeprazine, droperiodol and physeptone) and atropine. As we might expect, the regression analysis shows that dose is strongly associated with the probability of responding (T = 3.86, p = 0.0001). The estimated relationship is specified by the equation

$$\Pr(Y=1\,|\,d) = \frac{e^{-1.92+2.78d}}{1+e^{-1.92+2.78d}},$$

and figure 11.1 is actually a plot of this estimated dependence of $\Pr(Y = 1|d)$ on dose.

In many dose-response problems, the estimation of the median effective dose, which is denoted by ED_{50}, is of particular interest. The ED_{50} value is the actual dose required to induce the designated response in 50% of the population. Since $\Pr(Y = 1|d = ED_{50}) = \Pr(Y = 0|d = ED_{50}) = 0.5$, it follows that

$$\begin{aligned}\log\left\{\frac{\Pr(Y=1|d=ED_{50})}{\Pr(Y=0|d=ED_{50})}\right\} &= \log(0.5/0.5)\\ &= 0 = a + b(ED_{50});\end{aligned}$$

therefore $ED_{50} = -a/b$ and its estimated value is $\widehat{ED}_{50} = -\hat{a}/\hat{b}$. For the data analyzed in table 11.9 the estimated ED_{50} is $-(-1.92)/2.78 = 0.69$, which corresponds to a concentration of $\exp(0.69) = 1.99$ mg/kg. The derivation of a 95% confidence interval for this concentration is somewhat complicated, since the ED_{50} estimate is itself a ratio of estimates. Readers are advised to consult a statistician if a confidence interval of this type is required. In this particular case, most approaches to the technical problem of deriving a confidence interval for ED_{50} would yield an interval which is approximately (0.43, 0.95). The 95% confidence interval for the corresponding concentration is therefore $(e^{0.43}, e^{0.95}) = (1.54, 2.59)$. The derivation of estimates and confidence intervals for other doses or concentrations which are similarly defined is handled in a corresponding way.

To illustrate some special aspects of the use of logistic regression methods we now consider a combined analysis of the dose-response data for two of the patient groups described by Duncan et al. [20]. The first set of patients (Group 1) consists of 94 children who did not receive premedication, whereas the second set (Group 2) consists of the 137 children premedicated orally with TDP and atropine. The regression model which was fitted to these data involves three covariates, X_1, X_2 and X_3. The binary variable X_1 takes the value 0 if a patient did not receive premedication and 1 if the patient received TDP and atropine. The variable X_2 corresponds to the logarithm of the concentration of thiopentone administered, i.e., the dosage d, while the covariate X_3 represents an interaction term which is formed by multiplying X_1 and X_2. Thus, X_3 is equal to the dosage, d, for patients in Group 2 and takes the value 0 for patients in Group 1. This means that the fitted model consists of two different forms, one for each group of patients. These two forms are

$$\Pr(Y=1|d) = \begin{cases} \dfrac{\exp(a+b_2d)}{1+\exp(a+b_2d)} \text{ for Group 1,} \\ \\ \dfrac{\exp(a+b_1+b_2d+b_3d)}{1+\exp(a+b_1+b_2d+b_3d)} \text{ for Group 2.} \end{cases} \tag{11.3}$$

Table 11.10. Maximum likelihood fit of a binary logistic regression model to data on 231 children anaesthetized with thiopentone

Regression coefficient	Estimate	Estimated standard error	Test statistic
a	–5.59	1.32	–
b_1	3.67	1.56	2.35 ($p = 0.02$)
b_2	3.33	0.79	4.22 ($p < 0.000l$)
b_3	–0.55	1.06	0.52 ($p = 0.60$)

A total of 137 children were premedicated with TDP and atropine; the remaining 94 did not receive premedication.

Notice that a and b_2 in the model for Group 1 are replaced in the model for Group 2 by ($a + b_1$) and ($b_2 + b_3$), respectively. The regression coefficients b_1 and b_3 reflect two kinds of differences between the patient groups. To understand these differences, let $p_1(d)$ and $p_2(d)$ be the probabilities of responding to dose level d of the anaesthetic in Groups 1 and 2, respectively. Then the odds ratio (OR) for Group 2 versus Group 1 is

$$\frac{p_2(d)/\{1-p_2(d)\}}{p_1(d)/\{1-p_1(d)\}}=\frac{\exp(a+b_1+b_2d+b_3d)}{\exp(a+b_2d)}=\exp(b_1+b_3d).$$

If b_1 and b_3 are both zero, then this odds ratio is one and the probability of responding is identical in both groups. If $b_3 = 0$, then the odds ratio at all dose levels has the same value, namely $\exp(b_1)$. Thus, the dependence on dose is the same in both groups of patients and is described by the coefficient b_2. However, if $b_3 \neq 0$, then the dose dependence is different in the two groups, and therefore the odds ratio changes with dose by a factor $\exp(b_3d)$.

Table 11.10 presents the estimation of model (11.3). The regression coefficients for X_1 (Group 2) and X_2 (dose) are both significant, but the coefficient for X_3 is not significant. Thus, the dependence on dose of the probability of responding to thiopentone is the same in the two groups of patients. However, the actual probability of responding is higher for the group of patients who were premedicated with TDP and atropine. This difference between the two patient groups is reflected in the odds ratio, $\exp(b_1)$, which has an estimated value of $\exp(\hat{b}_1) = \exp(3.67) = 39.3$.

12

Regression Models for Count Data

12.1. Introduction

In some epidemiological or clinical studies, the response of interest consists of a count, such as the number of cells that show definite evidence of differentiation, or the number of repeated infections experienced by a subject. The values recorded will be only non-negative integers.

In some instances, it may be possible to analyze observed data that are counts using the methods of multiple linear regression that we described in chapter 10. However, regression methods are available that are better suited to response measurements that are counts, and we discuss the most commonly used method, which is known as Poisson regression, in this chapter.

Because it often provides a satisfactory representation for the variability observed in count data, the Poisson distribution plays a role in their analysis that is similar to that of the normal distribution in multiple linear regression, and the binomial distribution in logistic regression. The first occasion when the Poisson distribution was used to characterize observations that were counts appears to have occurred at the end of the 19th century, when Ladislaus von Bortkiewicz [21] showed that, over a 20-year period, the annual number of deaths attributed to horsekicks suffered by corpsmen in each of 14 Prussian army corps could be fitted very convincingly by a Poisson distribution. However, the name Poisson derives from a French mathematician, Siméon-Denis Poisson, who derived the mathematical form of the distribution.

A more recent, slightly unusual, medical example in which the Poisson distribution was used to summarize the variation in observed counts was in the analysis of a randomized trial, conducted by Fallowfield et al. [22], that was

Table 12.1. The results of a Poisson regression analysis of the number of focussed and/or open questions asked by a physician during a patient consultation

Explanatory variable	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
Course	0.24	0.05	5.06	<0.001
Physician sex	0.11	0.05	2.09	0.037
Seniority	–0.02	0.05	–0.45	0.651

designed to study the effect on physician communication skills of an intensive three-day training course.

Here, we ignore additional trial complexity, and consider a comparison of 80 doctors who were randomized to attend the three-day course with 80 doctors who were randomly chosen not to attend. We also restrict attention to a single outcome measure, namely the number of focussed and/or open questions asked by a physician during a patient consultation that occurred three months after the course ended, or three months after randomization for those physicians who did not receive any communication skills training. The course was designed to increase the frequency of such questions. For each physician, data were available from two consultations; to avoid undue complexity, we ignore the expected correlation between counts for the same physician and simply assume we have 160 observed counts for both the treatment group, i.e., those physicians who received training, and the control group.

Table 12.1 summarizes the results of a Poisson regression analysis of the number of focussed and/or open questions asked. The regression model included three explanatory variables that coded course attendance (yes = 1, no = 0), physician sex (female = 1, male = 0) and physician seniority (senior = 1, junior = 0). Readers will observe immediately that the format of this table is similar to those we first introduced in chapter 10 and subsequently encountered in chapter 11.

12.2. The Model for Poisson Regression

The theoretical formula from which we can calculate probabilities for counts that follow a Poisson probability function is characterized by a single parameter that is usually represented by the Greek letter λ. Conveniently, λ turns out to be the theoretical mean of the corresponding Poisson distribution, so that if we have an estimated value for λ, we can immediately calculate the corresponding probability that a count equal to y is observed in a Poisson distribution with mean λ. Since λ is the only adjustable parameter in this Poisson model for the variation in observed counts, it is natural to link λ to the values of explanatory variables of interest. Because the mean, λ, of a Poisson distribution must be greater than zero, it would be unsuitable simply to assume that

$$\lambda = a + b_1X_1 + \ldots + b_kX_k = a + \sum_{i=1}^{k} b_iX_i,$$

where $X_1, \ldots, X_k$ represent the values of various explanatory variables, such as coding for the sex of a physician. Unless we restrict the values of a and the regression coefficients $b_1, \ldots, b_k$, the right-hand side of this equation for λ could sometimes be a negative value.

The logarithmic transformation is a remedy for this dilemma; the sign and magnitude of log λ is completely unrestricted, making the logarithm of the Poisson mean a natural choice to equate to the expression, $a + \Sigma_{i=1}^{k} b_iX_i$, the component of the Poisson regression model which is the same as that which occurs in other regression models. Thus, if $X_1, \ldots, X_k$ are potential explanatory variables whose values we wish to use to model variability in a response measurement, Y, that is thought to follow a Poisson distribution, then using the equation

$$\log \lambda = a + b_1X_1 + \ldots + b_kX_k$$

is a natural way of allowing the measured values of these explanatory variables to account for the variability in observed values of Y.

As in other regression models that we have previously considered, if a particular regression coefficient, say b_i, is zero, then the corresponding explanatory variable, X_i is not associated with the response, Y. Thus, if there is no evidence to contradict the hypothesis that b_i equals 0, then we probably can omit X_i from a Poisson regression model for the observed data. As we discussed in §11.2 for the case of logistic regression, a suitable statistic for testing the hypothesis that the regression coefficient, b_i, equals zero is

$$T = \frac{|\hat{b}_i|}{\text{est. standard error}(\hat{b}_i)}.$$

The results of an analysis may also be presented in terms of the ratio

$$\frac{\hat{b}_i}{\text{est. standard error}(\hat{b}_i)},$$

which is equal to T, apart from the sign. The latter is the ratio found in table 12.1. Whichever version of this test statistic is used, the conclusion regarding the associated explanatory variable, X_i, is the same.

The explanatory variables used in fitting the Poisson regression model summarized in table 12.1 are all binary ones that encode whether or not a physician in the study attended the training course, was a female, and was more senior.

There is considerable evidence in the study data that the estimated regression coefficient associated with attending the course is significantly different from zero, establishing a behavioural effect that is associated with the training provided. There is also some evidence of an effect associated with a physician's sex, but there is no evidence of different behaviour patterns between senior and junior physicians. The signs of the estimated regression coefficients for course attendance and physician sex each indicate that the estimated Poisson mean is larger if the physician is female or if he or she attended the communication skills training.

The results of an analysis based on a Poisson regression model can also be described in terms of a rate ratio or 'relative rate'. If b_j is the regression coefficient associated with a particular binary explanatory variable, such as course attendance, $\exp(b_j)$ represents the ratio of the rate at which the events of interest occur among physicians who received the skills training compared to those who did not. Thus, the key feature of the analysis that we can distill from table 12.1 is that the rate at which physicians asked focussed and/or open questions, adjusted for sex and seniority, is $\exp(0.24) = 1.27$ times greater after attending the training course. And if we use the estimated standard error for $\hat{b}_1$ of 0.05 to derive the 95% confidence interval $0.24 \pm 1.96(0.05)$, i.e., (0.14, 0.34), for b_1, then a corresponding 95% confidence interval for the relative rate is $(\exp(0.14), \exp(0.34))$ or (1.15, 1.40).

Of course, the importance of an effect may be linked to its absolute rather than relative size. In this communication skills study, the relative rate effect of 1.26 was associated with a mean number of 6.54 focussed and/or open questions asked during a patient consultation by a physician in the trained group compared with a mean of 5.14 in the control group. Providing such information, in addition to the Poisson regression results listed in table 12.1, is sensible and informative.

As in the case of logistic regression, the calculations involved in fitting a Poisson regression model to observed data are known as maximum likelihood estimation, the details of which are beyond the intended scope of this book. Even though many software packages that are now available will fit Poisson regression models, there are some aspects of these models that may require careful attention in any particular analysis. Thus, readers may wish to consult a statistician when the use of Poisson regression seems appropriate. However, we hope that our brief introduction to this regression model for count data has been informative, and will enable readers to understand the use of this statistical methodology in published papers.

Although the preceding discussion is the standard approach to motivating, and characterizing, the Poisson regression model, there are at least two other ways in which a Poisson distribution might be thought appropriate for a given set of data, even when the response measurements are not obviously observed counts. This same technique is often used to analyze cohort studies when there are counts of events, such as deaths, but the length of total observation time for both the deaths and those who survive is important. In such a situation, cumulative exposure time is used as a denominator for the observed counts, e.g., the Standardized Mortality Ratio and Cumulative Mortality Figure. Interested readers can find a discussion of both these measures in Breslow and Day [23], and a related example is discussed in chapter 21. Likewise, a Poisson regression model can arise as the limiting approximation to what might otherwise constitute logistic regression with a very large sample size, n, and a very small probability, p, of observing the event of interest. This situation typically occurs when the probability that an individual subject experiences an outcome of interest is very small and, therefore, a substantial number of subjects are enrolled in the study so that a reasonable number of events can be observed during follow-up.

The example of Poisson regression that we consider in the following section represents yet another situation, and provides additional evidence that Poisson regression models are useful in a variety of different settings.

12.3. An Experimental Study of Cellular Differentiation

Trinchieri et al. [24] describe an experiment to investigate the immunoactivating ability of two agents – tumour necrosis factor (TNF) and immune interferon-γ (IFN) – to induce monocytic cell differentiation in human promyelocytes, which are precursor leukocytic cells. Following exposure to none, only one, or both agents, individual cells were classified as exhibiting, or lacking, markers of differentiation. The study design involved 16 different dose combinations of TNF and IFN; after exposure and subsequent incubation for five days, between 200 and 250 cells were individually examined and classified, although the precise numbers of cells involved were not available. Consequently, the observed data are simply counts of differentiated cells, rather than the observed numbers of both types of cells after exposure to one of the 16 dose combinations of the two agents.

If we knew the number of undifferentiated cells counted in each of the 16 experimental treatment combinations, we could consider using a logistic regression dose-response model to study the possible systematic relationship between the dose levels of TNF/IFN administered, and the observed proportion of differentiated cells. Since the undifferentiated cell totals were not available, we will use a Poisson regression model to investigate the same questions, treating the differentiated cell counts as the observed responses. In doing so, we are implicitly assuming that any variation in the total number of cells examined for each dose combination of TNF and IFN is unimportant.

As we have previously noted, dose-response studies such as this one often use a logarithmic concentration as the measurement scale of the experimental agent. In this instance, we chose to use base-10 logarithms, since the concentrations of TNF (in units per ml) used in the study design were 0, 1, 10 and 100, i.e., mostly ten-fold increases in concentration. To avoid numerical problems in fitting the Poisson regression model, we also replaced the absence of TNF, i.e., a concentration of 0 U/ml, by 0.1 U/ml, and the complete absence of IFN, i.e., 0 U/ml of IFN, by 0.1. Thus, the concentration levels 0, 1, 10 and 100 U/ml of TNF used in the experiment became the four doses –1, 0, 1 and 2 log U/ml, respectively. Likewise, the concentrations 0, 4, 20 and 100 U/ml of IFN used in the study corresponded to the four dose levels –1, 0.602, 1.301 and 2 log U/ml. For convenience, we will refer to these two dose measurements as the explanatory variables X_1 and X_2.

Through their study, Trinchieri et al. [24] sought to establish whether the two agents acted independently, or synergistically, to stimulate a differentiation response in cells. In the context of a Poisson regression model, this can be examined by creating a third explanatory variable, $X_3 = X_1 \times X_2$, which is called the interaction of X_1 and X_2. Another example of an interaction term was used in §11.5. If the regression coefficient, b_3, associated with X_3 is nonzero, then the effects of X_1 and X_2 cannot be treated separately in the sense that we would simply add the individual terms, b_1X_1 and b_2X_2, to the regression model. Instead, the event rates depend on the specific combination of X_1 and X_2 values because $X_3 = X_1 \times X_2$ is present in the model.

Accordingly, the equation for the mean of the Poisson regression model that we fitted to the number of differentiated cells observed for each of the 16 treatment combinations was

$$\log\{\lambda(X_1, X_2, X_3)\} = a + b_1X_1 + b_2X_2 + b_3X_3.$$

If the regression coefficient b_3 is equal to zero, changing the dose of IFN by one unit will have the same effect on the mean of the Poisson distribution, regardless of the dose of TNF used. However, if b_3 is different from zero, the effect on the mean of the Poisson distribution of a one-unit change in IFN will depend on how much TNF is present, and on the sign and value of b_3.

Table 12.2 summarizes the results of fitting this model involving the explanatory variables X_1, X_2 and X_3 to the observed data on cell differentiation.

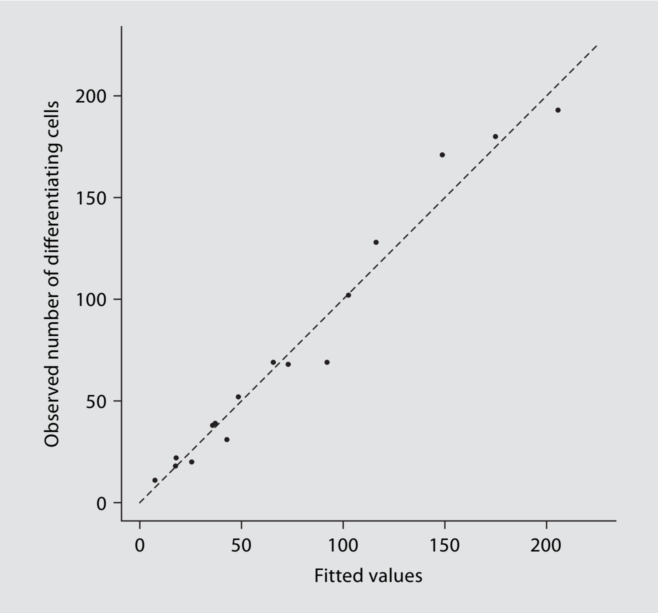


Fig. 12.1. Graphic display of the Poisson regression model $\log\{\lambda(X_1, X_2, X_3)\} = 3.32 + 0.773X_1 + 0.434X_2 - 0.101X_3$ fitted to the cell differentiation data; observed responses versus fitted values.

Table 12.2. The results of a Poisson regression analysis of the relationship between cellular differentiation and the dose of TNF or IFN administered

Regression coefficient	Estimate	Estimated standard error	Test statistic	Significance level (p-value)
a	3.32	0.079	–	–
b_1	0.773	0.048	16.0	$<10^{-15}$
b_2	0.434	0.052	8.3	$<10^{-15}$
b_3	−0.101	0.033	−3.1	<0.002

Although a thorough evaluation and interpretation of the results summarized in table 12.2 would be appropriate, we will simply note that all three explanatory variables are statistically important sources of systematic variation in the observed counts of the number of differentiating cells. Since the regression coefficient associated with the interaction term, X_3, is significantly different from 0, the differentiated cell counts represent evidence that the two agents, TNF and IFN-γ, do not independently stimulate cell differentiation; instead, the effect of either agent depends on the level of the other. In this particular model, a negative sign for the estimated value of b_3 means that the effect of a high concentration of one agent is reduced if the other agent is also present at a high concentration.

Figure 12.1 displays a plot of the observed numbers of differentiated cells, based on the fitted Poisson regression model, against the predicted counts from the experiments. This graphical display, as well as other plots that we have chosen not to reproduce here, indicate that the fitted model provides a very satisfactory statistical account of the apparent systematic dependence of the response measurement on the concentration of the two agents used in the study.

13

Proportional Hazards Regression

13.1. Introduction

In studies of chronic disease, perhaps the most frequent endpoint of interest is survival. In chapter 6, we introduced the Kaplan-Meier estimated survival curve and in chapter 7, the comparison of two survival curves via the log-rank test was described. In this chapter, we consider regression models for survival data. As with other regression models, the simultaneous influence of a number of explanatory variables can be investigated with these models. However, the models themselves have been designed to be particularly suited to survival data.

Throughout this section, we will refer to survival time as our endpoint of interest. As with chapter 6, however, the methodology applies to data arising as the time to a well-defined event. In cancer clinical trials, for example, the variable time to relapse or death is frequently used.

Two characteristics of survival data led to the development of regression models specific to this type of data. The first was the frequent occurrence of incomplete observation. If a number of patients are being followed to estimate the characteristics of a survival function, a fraction of these patients will very likely still be alive at the time of analysis. Therefore, we do not have complete survival information on these patients. Nevertheless, the knowledge that they have survived for a certain period of time should be incorporated into the analysis.

The second characteristic of survival data that led to the development of new statistical methods is the distribution of typical survival times. Classical statistical theory, such as multiple linear regression, is usually not appropriate for clinical data. In most studies, the distribution of survival times is unknown and the distributions tend to vary widely from one disease to another. Either the statistical models must incorporate a wide class of distributions that may be less readily applied than the normal distribution, or we must use methods that do not assume a specific distribution at all. Our discussion of proportional hazards regression concentrates on methodology that adopts the latter approach.

In chapters 6 and 7, we discussed the survival experience of 64 patients with Stages II, III or IV non-Hodgkin's lymphoma. A comparison was made between those whose disease presented with clinical symptoms (B symptoms) and those whose disease was discovered indirectly (A symptoms). A strong survival advantage was attributed to those patients with A symptoms.

A natural question to ask is whether this difference could arise because these two patient groups differed with respect to other important prognostic factors. For illustration, we will consider two such factors, stage of disease and the presence of a bulky abdominal mass.

13.2. A Statistical Model for the Death Rate

When analyzing survival time, it is convenient to think in terms of the death rate. More generally, a failure rate can be defined for investigating the time to any other type of 'failure'. The technical name for a failure rate is the hazard rate; historically, the death rate was called the force of mortality.

The death rate is defined, in mathematical terms, as the probability of dying at a specified time t when it is known that the individual did not die before t. This definition may seem somewhat artificial. However, it is a convenient, general way of representing the same type of information as that implied by the question 'What is the probability of surviving two years after therapy if a patient has already survived one year?'.

We will denote a death rate at time t by the function d(t). To develop a regression model, we need to incorporate information which is coded as covariates $X_1, X_2, ..., X_k$. As before, it is convenient to refer to the set of covariates as $\underset{\sim}{X} = \{X_1, X_2, ..., X_k\}$. Thus, in the regression model we define, we want to describe the quantity $d(t; \underset{\sim}{x})$, which is the death rate for an individual with observed values of the covariates $X_1 = x_1, X_2 = x_2, ..., X_k = x_k$, or $\underset{\sim}{x} = \{x_1, x_2, ..., x_k\}$. A convenient regression model for $d(t; \underset{\sim}{x})$ is specified by the equation

$$\log\{d(t; \underset{\sim}{x})\} = \log\{d_0(t)\} + \sum_{i=1}^{k} b_i x_i. \tag{13.1}$$

The death rate function, $d_0(t)$, in equation (13.1) represents the death rate at time t for an individual whose covariate values are all zero, i.e., $X_1 = 0$, $X_2 =$

$0, ..., X_k = 0$. It is not important that a patient having all values of the covariates equal to zero be realistic. Rather, $d_0(t)$ is simply a reference point, and serves the same function as the coefficient a in the expression $a + \Sigma b_i x_i$ which we used in previous regression models. The difference which $d_0(t)$ incorporates is that $d_0(t)$ changes as t varies, whereas the coefficient a was a single numerical constant. This model was introduced by Cox [25] and is frequently referred to as the Cox regression model.

A survival curve is a graph of the function $\Pr(T > t)$, the probability of survival beyond time t. For those who remember some calculus, we remark that

$$\Pr(T > t) = \exp\left\{-\int_0^t d(y)dy\right\}.$$

Otherwise, it suffices to say that by specifying the death rate, d(t), we identify the function $\Pr(T > t)$. Moreover, in terms of $\Pr(T > t)$, it can be shown that the regression equation (13.1) is equivalent to

$$\Pr(T > t; \underset{\sim}{x}) = \left\{\Pr_0(T > t)\right\}^{\exp\left\{\sum\limits_{i=1}^{k} b_i x_i\right\}}, \tag{13.2}$$

where $\Pr_0$ $(T > t)$ is the survival curve for a patient whose covariate values are all zero.

Once again, the most important aspect of the regression model (13.1) concerns the regression coefficients. If b_i is zero, then the associated covariate is not related to survival, or does not contain information on survival, when adjustment is made for the other covariates included in the model.

The model is also valuable, however, because it easily accommodates the two specific characteristics of survival data which we mentioned in §13.1. When we estimate the coefficients, the b_i's, it is easy, and is in fact preferable, not to assume anything about the death rate function $d_0(t)$. Thus, we do not need to specify any particular form for the distribution of the survival times, and the model is applicable in a wide variety of settings. Secondly, it turns out that it is very easy to incorporate the information that an individual has survived beyond time t, but the exact survival time is unknown, into the estimation of the regression coefficients. The details of the actual calculations are beyond the scope of this book but, as before, the model estimation can be summarized as a table of estimated regression coefficients and their corresponding estimated standard errors.

Table 13.1. The results of a proportional hazards regression analysis of survival time based on 64 advanced stage non-Hodgkin's lymphoma patients

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic
Stage IV disease	1.38	0.55	2.51 (p = 0.012)
B symptoms	1.10	0.41	2.68 (p = 0.007)
Bulky disease	1.74	0.69	2.52 (p = 0.012)

13.3. The Lymphoma Example

Table 13.1 records the estimated regression coefficients and standard errors from a Cox regression analysis of survival time for the 64 lymphoma patients described in chapter 7. Three covariates, X_1, X_2 and X_3 were defined as follows:

$$X_1 = \begin{cases} 1 & \text{if the disease is Stage IV,} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the patient presents with B symptoms,} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if a large abdominal mass} \, (> 10\,\text{cm}) \text{ is present,} \\ 0 & \text{otherwise.} \end{cases}$$

Coincidentally, these covariates are all binary, although this is not a formal requirement of the proportional hazards regression model.

From table 13.1, we see that all three covariates are related to survival, even after adjustment for the other covariates. Thus, for example, we can say that the prognostic importance of B symptoms is not due to differences in stage and abdominal mass status among patients in the two symptom groups.

With a proportional hazards or Cox regression model and binary covariates, the results of an analysis can easily be described in terms of relative risk. For example, the death rate for Stage IV patients is $\exp(\hat{b}_1) = \exp(1.38) = 3.97$ times as large as that for Stage II or III patients. The relative risk associated with Stage IV disease is therefore 3.97. The relative risks for B symptoms and bulky disease are $\exp(1.10) = 3.00$ and $\exp(1.74) = 5.70$, respectively.

In this regression model, the relative risks need to be calculated relative to a baseline set of patient characteristics. For our model with three binary covariates, it is convenient to take $X_1 = 0$, $X_2 = 0$ and $X_3 = 0$ as the baseline characteristics. Then the estimated relative risk for an individual with $X_1 = x_1$, $X_2 = x_2$ and $X_3 = x_3$ is

$$\exp\left(\hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3\right). \tag{13.3}$$

For example, the relative risk for a Stage IV patient with B symptoms and bulky disease is exp(1.38 + 1.10 + 1.74) = 68.0 compared to a Stage II or III patient with A symptoms and no abdominal mass. Estimated relative risks for such extreme comparisons should be interpreted with some skepticism, since the associated standard errors are frequently very large. It is also possible to examine whether the model specified in equation (13.3), which sums the three regression components, is, in fact, reasonable. To do this, we would code additional interaction covariates which represent patients having two or three of the characteristics of interest. If these extra covariates have non-zero regression coefficients, then they should be included in a revised model, and their inclusion will alter the additive relationship in equation (13.3).

For example, if we code a covariate X_4 as

$$X_4 = X_1 \times X_2 = \begin{cases} 1 & \text{if a patient has B symptoms and Stage IV} \\ 0 & \text{otherwise,} \end{cases}$$

then the estimated regression model including X_1, X_2, X_3 and X_4 has regression coefficients $\hat{b}_1 = 2.17$, $\hat{b}_2 = 1.98$, $\hat{b}_3 = 1.66$ and $\hat{b}_4 = -1.07$. The coefficient for X_4 has an associated p-value of 0.36. From this model, the estimated relative risk for a Stage IV patient with B symptoms and bulky disease is exp(2.17 + 1.98 + 1.66 – 1.07) = 114.4. This estimate is presented solely for the purposes of illustration, since the p-value of 0.36, which is associated with a test of the hypothesis $b_4 = 0$, does not support the use of a model incorporating X_4. In fact, with a small data set, a model involving a covariate which represents several factors combined usually would not be considered.

The recognition that regression coefficients are only estimates is more important than the particular estimates of relative risk which can be calculated. As we indicated in chapter 8, we can calculate confidence intervals for regression coefficients. The 95% confidence interval for b_i is defined to be

$$\hat{b}_i \pm 1.96\{\text{est. standard error } (\hat{b}_i)\}.$$

If we represent this interval by (b_{iL}, b_{iH}), then, for a Cox regression model, we can also obtain a 95% confidence interval for the relative risk, $\exp(b_i)$, namely

$$(\exp(b_{iL}), \exp(b_{iH}))\cdot$$

Table 13.2 presents estimated regression coefficients, relative risks and confidence intervals for the lymphoma example. Since there are only 64 patients,

Table 13.2. Estimates of the 95% confidence intervals for regression coefficients and relative risks corresponding to the model analyzed in table 13.1

Covariate	Regression coefficient	95% Confidence interval	Relative risk	95% Confidence interval
Stage IV disease	1.38	(0.30, 2.46)	3.97	(1.35, 11.68)
B symptoms	1.10	(0.30, 1.90)	3.00	(1.35, 6.71)
Bulky disease	1.74	(0.39, 3.09)	5.70	(1.47, 22.03)

the confidence intervals are very wide; therefore, little importance should be attached to any of the estimates. Any discussion of estimation based on regression models should clearly indicate the potential error in the estimates. Remember, also, that a 'significant' regression coefficient ($p < 0.05$) merely means that a 95% confidence interval for the regression coefficient excludes the value zero.

It is not possible to discuss, in one brief chapter, all the intricacies of a proportional hazards regression model. We would suggest that a statistician be involved if this model is used extensively in the analysis of any data set. However, there is one remaining feature of the model which merits some discussion.

Any regression model is based on a particular, assumed relationship between the dependent variable and the explanatory covariates. It is important to examine the validity of this assumed relationship. It need not have a biological basis, but it must be a reasonable, empirical description of the observed data. Most regression models allow the model assumptions to be examined.

The assumption of the Cox regression model which is discussed most often is that of proportional hazards. Equation (13.1) implies that

$$d(t; \underset{\sim}{x}) = \{d_0(t)\} \times \exp\left\{\sum_{i=1}^{k} b_i x_i\right\},$$

which specifies that the death rate for an individual with covariate values $\underset{\sim}{x}$ is a constant multiple, $\exp\{\Sigma\ b_i x_i\}$, of the baseline death rate *at all times.* Thus, although the death rate $d_0(t)$ can change with time, the ratio of the death rates, $d(t; \underset{\sim}{x})/d_o(t)$, is always equal to $\exp\{\Sigma\ b_i x_i\}$.

The proportional hazards regression model is frequently criticized for this assumption but, in fact, it is very easy to generalize the model to accommodate a time-dependent hazard or death rate ratio. In the following discussion, we indicate a simple approach which is useful in many situations.

Let us suppose that we do not want to assume that the death rate for Stage IV patients is a constant multiple of that for Stage II or III patients, and we do

Table 13.3. The results of a proportional hazards regression analysis of survival time, stratified by stage of disease; the analysis is based on 64 advanced stage non-Hodgkin's lymphoma patients

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic
B symptoms	1.11	0.41	2.71 ($p = 0.007$)
Bulky disease	1.80	0.69	2.61 ($p = 0.010$)

not have any information concerning the actual nature of the relationship. In this case, two regression equations can be defined as follows:

$$\log\{d_1(t; x_2, x_3)\} = \log\{d_{01}(t)\} + b_2x_2 + b_3x_3$$

and

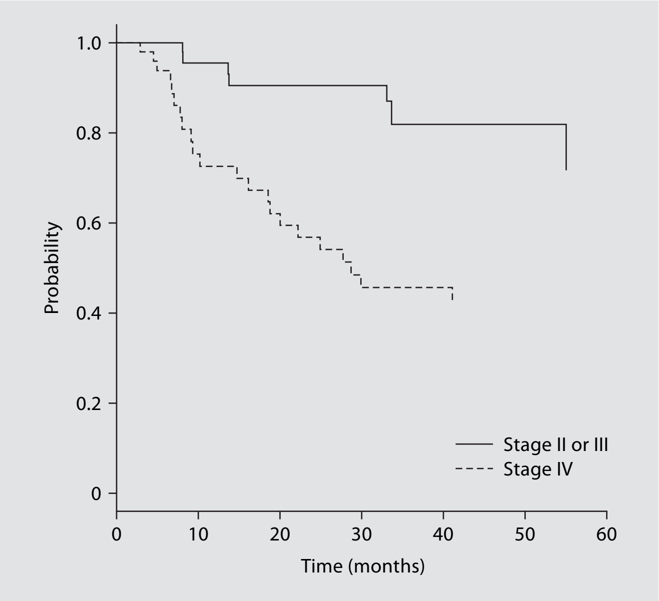
$$\log\{d_2(t; x_2, x_3)\} = \log\{d_{02}(t)\} + b_2x_2 + b_3x_3,$$

where $d_1(t; x_2, x_3)$ is the death rate for Stage II or III patients and $d_2(t; x_2, x_3)$ is the death rate for Stage IV patients. Notice that the regression coefficients for B symptoms and bulky disease are assumed to be the same in the two regression equations. This is called a stratified version of Cox's regression model, and it is an extremely useful generalization of the basic proportional hazards regression model.

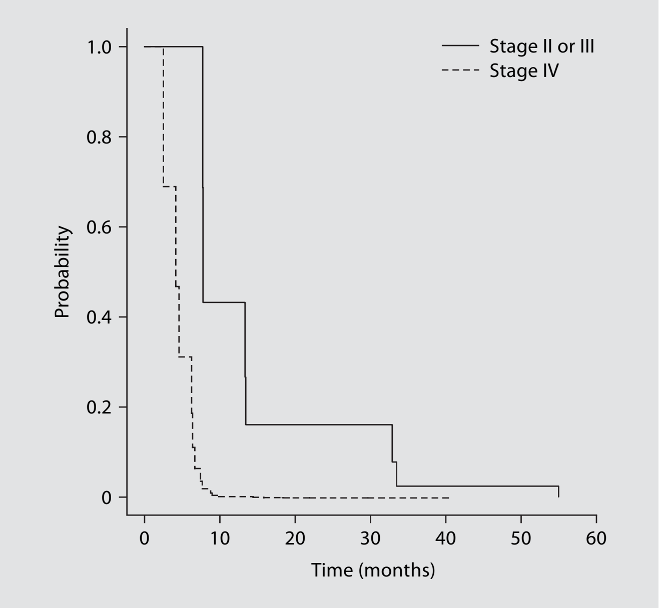
Table 13.3 presents estimates and standard errors of b_2 and b_3 based on this stratified model. In table 13.1, the test that b_2 or b_3 could be equal to zero was adjusted for any possible confounding effects of disease stage by the inclusion of the covariate X_1 in the regression model. In table 13.3, the corresponding tests are adjusted because the model is stratified by disease stage. This latter adjustment is more general, in that it does not assume there is any specific relationship between $d_1(t; x_2, x_3)$ and $d_2(t; x_2, x_3)$, whereas, in table 13.1, the assumption has been made that

$$d_2(t; x_2, x_3) = d_1(t; x_2, x_3) \exp(b_1).$$

In this example, the conclusions with respect to b_2 and b_3 are hardly altered by the use of the stratified model. This will not always be the case, of course. When a stratified model is used, it is possible to estimate the baseline functions $d_{01}(t)$ and $d_{02}(t)$. This means, in our particular example, that we can estimate survival curves for patients with A symptoms and no bulky disease, subdivided by disease stage (II or III versus IV). Figure 13.1 presents these estimates graphically. The method of estimation will not be discussed here. It is a generalization of the techniques which we discussed in chapter 6. If we then



1



2

Fig. 13.1. The estimated survival curves for non-Hodgkin's lymphoma patients with A symptoms and no bulky disease, stratified by disease stage.

Fig. 13.2. The estimated survival curves for non-Hodgkin's lymphoma patients with B symptoms and bulky disease, stratified by disease stage.

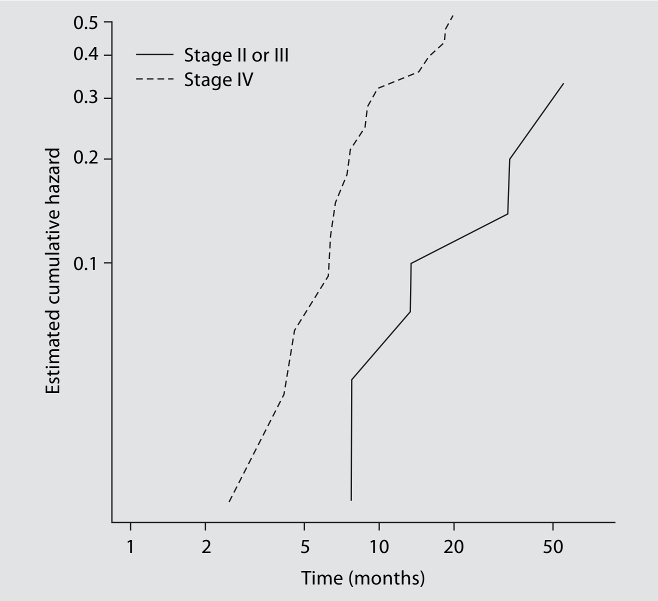


Fig. 13.3. The estimated cumulative hazard functions for the estimated survival curves shown in Figure 13.1, plotted using logarithmic scaling on both axes.

multiply the estimates of $d_{01}(t)$ and $d_{02}(t)$ by $\exp\{\hat{b}_2x_2 + \hat{b}_3x_3\}$, we can generate the corresponding survival curves for patients with covariate values $X_2 = x_2$ and $X_3 = x_3$. For example, figure 13.2 presents the curves for patients with B symptoms and bulky disease.

If we wish to examine whether a proportional hazard assumption is reasonable for representing the effect of disease stage (II or III versus IV), then figure 13.1 can be replotted, displaying the estimated cumulative hazard (–1 times the logarithm of the survival probability) versus time and using logarithmic scales on both axes. This is done in figure 13.3. If the two estimated survival curves are parallel, then the ratio of the death rates for the two patient groups is constant over time. In figure 13.3, the two curves are roughly parallel; therefore, it seems appropriate to include X_1 in an unstratified regression model.

13.4. The Use of Time-Dependent Covariates

The lymphoma example which was discussed in the previous section introduces most of the basic features of proportional hazards regression, with one notable exception. All the explanatory variables in the regression model for the death rate represent patient characteristics which are defined at the start of the follow-up period, and this initial classification does not change throughout the analysis. Covariates of this type are said to be fixed with respect to time, and commonly arise in clinical studies. However, in some situations it may be desirable, and appropriate, to examine the influence on a hazard rate of patient characteristics which change over time. In the remainder of this section, we describe such a situation and illustrate the ease with which time-dependent covariates can be incorporated into proportional hazards regression. In our view, it is a very attractive feature of this regression approach to the analysis of survival data.

Following bone marrow transplantation for the treatment of acute leukemia, an important outcome event is the recurrence of the disease. The rate of leukemic relapse can be modelled using a proportional hazards regression of the time from transplantation to leukemic relapse. Another serious complication which may arise in the immediate post-transplant period is acute graft-versus-host disease (GVHD), which is thought to be an immunologic reaction of the new marrow graft against the patient. The interrelationship of these two adverse outcomes is of particular interest.

Prentice et al. [26] examine this interrelationship by incorporating information on the occurrence of GVHD in a proportional hazards regression model for leukemic relapse following bone marrow transplantation. However, the development of GVHD in a patient is not a predictable phenomenon. Therefore, it would be quite inappropriate to model the effect of GVHD on the relapse rate by using a covariate which ignores this fact, i.e., by using a fixed covariate which classifies an individual as having GVHD throughout the post-transplant period. The simplest possible way to incorporate this temporal dependence would involve the use of a binary covariate which is equal to zero at times prior to a diagnosis of GVHD, but takes the value one at all times thereafter. If more comprehensive data are available, perhaps indicating the severity of GVHD, this information could also be incorporated into the regression model through the use of suitably defined time-dependent covariates.

In §13.2, we used the notation, d(t), to represent the dependence of a death rate on time. Here we will denote a relapse rate by r(t) and, in a similar fashion, use X(t) to show the dependence of a covariate on time. Let $\underset{\sim}{X}(t) = \{X_1(t), ..., X_k(t)\}$ represent the set of covariates in the regression model. Then for an individual with observed values of the covariates $\underset{\sim}{x}(t) = \{x_1(t), ..., x_k(t)\}$, a regression equation for the leukemic relapse rate, which parallels equation (13.1), is

$$\log\{r(t; \underset{\sim}{x}(t))\} = \log\{r_0(t)\} + \sum_{i=1}^{k} b_i x_i(t).$$

Table 13.4. The results of a proportional hazards regression analysis of leukemia relapse data based on 135 patients treated for acute leukemia by means of bone marrow transplantation

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic
GVHD	−0.76	0.37	2.05 (p = 0.04)
Transplant type	0.05	0.34	0.15 (p = 0.88)
Age	0.13	0.10	1.30 (p = 0.19)

Adapted from Prentice et al. [26] with permission from the publisher.

The addition of time-dependent covariates to the model is a very natural extension of equation (13.1), which already included a dependence on time. Although this refinement of the proportional hazards model appears to be simply a change in notation, it represents a major advance in biostatistical technique. There are many subtleties associated with its use and interpretation which we are not able to discuss adequately in these few pages. Readers are strongly urged to consult a statistician from the very beginning of any proposed study that may eventually involve the use of time-dependent covariates in a regression model.

Table 13.4, which is taken from Prentice et al. [26], presents the results of an analysis of data on leukemic relapse in 135 patients. Since the sample includes 31 syngeneic (identical twin) bone marrow transplants with no risk of GVHD, the regression model includes a binary covariate indicating the type of transplant (0 = syngeneic, 1 = allogeneic) and a continuous covariate representing patient age (years/10). As the results of this analysis indicate, neither of these covariates is significantly associated with leukemic relapse. However, their inclusion in the model adjusts the estimation of the GVHD effect for the influence of transplant type and patient age. Even after adjusting for the effect of these variables, the regression coefficient for GVHD is significant at the 0.05 level. The rate of leukemic relapse for patients who develop GVHD is estimated to be $\exp(-0.76) = 0.47$ times the relapse rate for patients who do not have GVHD at the same time post-transplant. The corresponding 95% confidence interval for this relative risk is $(e^{-1.46}, e^{-0.03}) = (0.23, 0.97)$.

The results of this analysis suggest that the occurrence of GVHD is protective with respect to leukemic relapse. This may indicate that GVHD serves to eradicate residual or new leukemic cells. The clinical implications of this finding are, of course, subtle and will not be discussed here. However, it suggests that although severe GVHD is clearly undesirable, a limited graft-versus-host reaction could help to control leukemic relapse.

14

The Analysis of Longitudinal Data

14.1. Introduction

In nearly all the examples that we have discussed in previous chapters, each study subject has yielded one value for the outcome variable of interest. However, many medical studies involve long-term monitoring of participants; therefore, the repeated measurement of an outcome variable is both feasible and likely. Sometimes, it may be reasonable to focus on one particular value in a series of measurements. More often, the full set of outcome variables measured will be of interest.

A variety of statistical methods have been developed for analyzing data of this type, i.e., longitudinal data. The essential difference between the various methods of analysis that we have discussed in previous chapters and the approach required for longitudinal data is that the model must account for the correlation between repeated observations on the same subject. That is, two observations on the same individual will tend to be more similar than two individual measurements taken on distinct subjects.

The classic example of a statistical method for analyzing studies involving more than one measurement on each subject is known as repeated measures analysis of variance. This topic is introduced in chapter 15, where we also provide additional details concerning the analysis of variance. However, that material is more technical than most subjects that we address in this book, so, in this chapter, we will avoid any discussion of analysis of variance. Instead, we will discuss three examples of longitudinal studies that allow us to illustrate some recently developed, quite general methods of analyzing longitudinal data.

14.2. Liang-Zeger Regression Models

14.2.1. The Study

The first study to be discussed is one concerning the relationship between the use of recreational drugs during sexual activity and high-risk sexual behavior in a cohort of 249 homosexual and bisexual men during a five-year period. The cohort was monitored approximately every three months and, at each follow-up visit, participants were interviewed in private by a trained interviewer.

For the purposes of illustration, we will describe only a simplified analysis of this study; readers who are interested in a more comprehensive discussion should consult Calzavara et al. [27]. Based on the interview conducted at a follow-up visit, each study subject was assigned a summary sexual activity risk score which we shall denote as RS. This score was designed to summarize both the risk level of the sexual activities in which the subject had participated during the previous three months and the number of partners with whom these activities had been performed. The average RS value across all subjects declined from a high of 152.2 on the first follow-up visit to a low of 60.0 on the 17th time the cohort was monitored. A logarithmic transformation of the RS observations was used to make the distribution of this outcome variable at each monitoring occasion similar to a normal distribution. In the remainder of this section, we will denote the logarithm of the RS measurement by the variable Y.

Although various explanatory variables were investigated in this study, the analysis that we describe below will use only three. The first, X_1, is an ordinal variable denoting the sequence number of the follow-up visit; the values of X_1 range from 1 to 20. The second, X_2, is a binary variable indicating that the study participant had used recreational drugs in conjunction with a sexual encounter in the previous three months ($X_2 = 1$); otherwise, $X_2 = 0$. The final explanatory variable, X_3, identifies whether the subject was HIV-1 seropositive at the preceding follow-up visit ($X_3 = 1$) or HIV-1 seronegative ($X_3 = 0$).

14.2.2. The Regression Model

It is natural to adopt a regression model to study the relationship between Y, the logarithm of RS, and the three explanatory variables. Symbolically, the equation

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

for this regression model is similar to the one we used in chapter 10 to describe the relationship between brain weight, body weight and litter size in preweaning mouse pups. However, the analysis in chapter 10 was based on the assumption that each value of Y measured was independent of all other values. Since the brain weight measurements used in chapter 10 were obtained from different litters, this independence assumption seems appropriate. In the present study, the same assumption of independence is unreasonable since a single subject may contribute up to 20 values of Y.

In recent years, Liang and Zeger [28, 29] have developed a method of analyzing longitudinal data using regression models. Their approach is based on an assumption about the correlation between observations on the same subject. A discussion of the range of possibilities that might be adopted in analyzing longitudinal data is beyond the scope of this book. The simplest assumption, and the one that we will adopt in analyzing the present study, is that the correlation between pairs of observations from the same subject does not vary between pairs; we will denote the unknown value of this common correlation by the Greek letter ρ. However, pairs of Y values that were obtained from different individuals are still assumed to be independent, which means their correlation is zero. In order to estimate the coefficients a, b_1, b_2, and b_3, in the regression model, it is necessary to incorporate the additional parameter ρ in the estimation procedure.

The Liang-Zeger approach to analyzing data from longitudinal studies involves two notable advantages. First, the method can be used with many different types of regression models. For example, if the response or outcome variable is binary and we wish to use a logistic regression model, the method of analyzing the data is essentially unchanged. The second advantage derives from the method of estimation, which is called generalized estimating equations (GEE); this methodology is often referred to as GEE regression models. According to statistical theory, the estimated regression coefficients are valid even if the correlation assumptions on which the analysis is based are not precisely correct. In most situations involving longitudinal data, the critical component in a sensible analysis is to incorporate some assumption about correlation so that the unreasonable premise that repeated measurements on the same subject are independent can be avoided. The Liang-Zeger approach confers the additional advantage that the 'robust' method of estimation even accommodates uncertainty about the most appropriate assumption concerning correlation.

One consequence of this 'robust' estimation procedure is that while the value of ρ is estimated, there is usually no corresponding estimated standard error. However, the importance of the correlation parameter is, in most instances, secondary, since the analysis primarily concerns the relationship between the outcome variable, Y, and the associated explanatory variables.

Table 14.1. The results of a Liang-Zeger regression analysis of longitudinal RS data obtained from a cohort of 249 homosexual and bisexual men

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
a	3.736	0.120	–	–
Visit number	–0.054	0.007	7.71	<0.001
Drug use	0.550	0.085	6.47	<0.001
Seropositive	0.443	0.254	1.74	0.081

$\rho = 0.544$.

14.2.3. Illustrative Results

Table 14.1 summarizes the results of fitting the regression model outlined in the previous section to the RS data.

The tabulated values are interpreted in the same way that estimated regression coefficients were understood in previous chapters involving regression models. A regression coefficient that is significantly different from zero represents an association between the outcome variable and the corresponding covariate, after adjusting for all other covariates included in the model. A statistical test of the hypothesis that the regression coefficient is zero can be based on the magnitude of the ratio of the estimated coefficient to its standard error; this ratio is compared to critical values from the modulus of a normal distribution with mean zero and variance one (cf. table 8.1).

According to the results presented in table 14.1, there is a demonstrable relationship between the visit number and the RS measurement; the negative sign of the estimated regression coefficient indicates that the mean of Y tended to decline as the study progressed. Since the regression coefficient associated with X_2 is positive and significantly different from zero, the use of recreational drugs during sexual encounters is associated with an increase in the mean value of Y. Finally, although there is an estimated increase in the mean response associated with HIV-1 seropositive status, these data provide no evidence that the increase is significantly different from zero. Since Y denotes the logarithm of the RS value, the corresponding conclusions with respect to the original RS measurement are that the mean value declined substantially as the study progressed; however, whenever recreational drugs were used during sexual encounters, the mean RS value at the succeeding interview tended to be higher.

Based on the results of this simplified analysis, the findings of this longitudinal study would appear to be that, on average, high-risk sexual activity in the cohort has declined over time. Nonetheless, high-risk activities, when they occur, tend to involve recreational drug use with a sexual encounter.

14.3. Random Effects Models

14.3.1. The Study

In a clinical investigation of ovulation, diabetic and healthy women were followed for various periods of time during which each ovulatory cycle that a subject experienced was classified as abnormal or normal. The hypothesis of interest was whether diabetic women had a higher frequency of anovulatory cycles than non-diabetic women.

The study involved 23 diabetic women and 58 who were not diabetic. The number of cycles classified for each woman varied from 1 to 12. Thus, this investigation represents an example of a data set consisting of relatively short sequences of binary data observed on a moderately large number of women. Since the character of cycles in the same woman should be similar, each observed cycle cannot be regarded as an independent observation. Therefore, as we indicated in our discussion of the previous example in this chapter, our analysis of the study data should take account of the correlation between ovulatory cycles of the same woman.

14.3.2. The Regression Model

Liang-Zeger regression models are often characterized as *marginal* regression models because the regression model itself looks just like one that might be used if only a single response measurement was available for each subject. Therefore, in some sense, it represents a model for any randomly selected observation from the population. In addition to being referred to as a marginal regression model, a Liang-Zeger regression model is also sometimes called a population-averaged model.

For binary data, the Liang-Zeger approach would use a logistic regression model. To analyze the study of ovulatory cycles in diabetic and healthy women, such a model could take the form

$$\Pr(Y = 1 | x) = \frac{\exp(a + bx)}{1 + \exp(a + bx)} \tag{14.1}$$

that we encountered in chapter 11, where Y = 1 denotes an abnormal cycle and Y = 0 a normal cycle. The explanatory variable X can be used to denote diabetic status with X = 1 corresponding to a diabetic woman and X = 0 otherwise. Then the associated regression coefficient, b, is the logarithmic odds ratio of an abnormal ovulatory cycle, and exp(b) is the corresponding odds ratio, reflecting the effect that being diabetic has on the probability of experiencing an abnormal cycle.

An alternative to the Liang-Zeger approach to dealing with two or more correlated observations from a woman in the study is to use a regression model that is similar to the stratified logistic regression model that we introduced in §11.3; see equation (11.2). This alternative regression model is represented by the equation

$$\Pr(Y=1|x)=\frac{\exp(a_i+bx)}{1+\exp(a_i+bx)} \tag{14.2}$$

where the subscript i indexes all the women in the study. By adopting a distinct value, a_i, of the intercept for each woman, this model specifies that the overall rate of an abnormal ovulatory cycle can vary arbitrarily among women. The model also assumes that this variation in the values of a_i can account for the correlation between observations that were obtained from the same woman. Note, however, that the effect of being diabetic on the probability of having an abnormal ovulatory cycle, which is measured by b, is assumed to be the same for all women in the study.

With many women and small numbers of observations from some subjects, it is not possible to estimate the large number of subject-specific parameters, i.e., the 81 values of a_i. However, if we also assume that the various a_i values all come from a common probability distribution, such as a normal distribution with a population mean and standard deviation denoted by a and σ, respectively, then we can estimate these latter two values.

In general, fitting such a random effects model can be quite a complex task, one that we choose not to discuss here. However, the simplest output from appropriate software will be similar to that which we have encountered for other regression models. Also, as we noted in the case of Liang-Zeger models, the general structure of such a random effects assumption can be incorporated into many different types of regression models, including those we have considered in the preceding four chapters.

The assumption that the subject-specific parameters a_i belong to a common probability distribution is why such a model is called a ‘random effects’ model. Notice, however, that the regression coefficient associated with the explanatory variable of interest, which denotes diabetic status in this example, measures how the odds in favour of an abnormal ovulatory cycle for any particular woman, with her specific value of a_i, would change if the woman was diabetic compared to the corresponding odds if she was not diabetic. Thus, in contrast to Liang-Zeger *marginal* models, this type of random effects model is sometimes called a *subject-specific* regression model. A full discussion of the various distinctions between these models is beyond the scope of this book,

Table 14.2. The results of two logistic regression analyses of longitudinal data collected from 81 women concerning diabetic status and abnormal ovulatory cycles

Regression coefficient	Estimate	Estimated standard error	Significance level
Ordinary logistic regression			
a	−0.72	0.15	–
b	0.55	0.26	0.032
Random effects logistic regression			
a[1]	−0.89	0.21	–
b	0.67	0.38	0.079

[1] Population mean.

but readers may encounter this terminology in other settings, and we hope our brief discussion has provided some useful background.

14.3.3. Illustrative Results

Among the 23 diabetic women, 43 of 106 ovulatory cycles (41%) were observed to be abnormal, while in the 58 healthy women, 51 of 181 cycles (28%) were abnormal. Table 14.2 summarizes the results of fitting two different logistic regression models to these data. The first analysis uses a simple logistic model that corresponds to equation (14.1), and the second is based on the random effects model specified in equation (14.2).

If we compare the results of the two different fitted models summarized in table 14.2, we see that in the ordinary logistic regression analysis, in which each of the 287 ovulatory cycles are treated as independent observations, the estimated regression coefficient associated with diabetic status is found to be significantly different from zero. However, although the corresponding estimated regression coefficient in the random effects model is roughly the same size, and has the same sign as its counterpart in the other analysis, the estimated standard error is larger in the random effects model and hence we would conclude that the data do not represent evidence to contradict the hypothesis $b = 0$. This outcome reflects a typical pattern that unfolds in such cases, namely that an analysis which fails to account for the correlation in longitudinal data appropriately is more likely to identify significant explanatory variable effects than one which does make some allowance for the correlation.

In table 14.2 the estimated mean of the subject-specific random effects is similar to the single estimated intercept, $\hat{a} = -0.72$, in the ordinary logistic regression. However, the two estimated values do not have the same interpretation. In addition, the estimated value of σ, the standard deviation of the common distribution of the subject-specific intercepts, is 1.02; this value is not the same as the tabulated standard error associated with the estimated mean of the distribution, i.e., 0.21.

If we base our conclusions concerning this study on the more appropriate logistic regression model that involves a random, subject-specific intercept, we can estimate that the odds of an abnormal ovulatory cycle are $\exp(0.67) = 1.95$ times greater for a diabetic woman than for a healthy woman. The corresponding 95% confidence interval for this odds ratio is $\exp\{0.67 \pm 1.96(0.38)\} = (0.93, 4.12)$, which includes 1. Therefore, in this limited data set the statistical evidence for an effect that being diabetic has on a woman's ovulatory cycles is marginal. Moreover, the estimated value of σ suggests that there is considerable variation from woman to woman with respect to the probability of experiencing an abnormal ovulatory cycle.

14.3.4. Comments

In the preceding two sections, we have provided a rather brief introduction to two relatively new statistical methods that use regression models to analyze data from longitudinal studies. We hope that the examples we have discussed in §§14.2 and 14.3 will provide a basis for understanding the presentation of such analyses in the medical literature. Readers who are interested in using either Liang-Zeger or random effects regression models to analyze a particular study should consult a statistician.

Dependence between observations is both a central feature of longitudinal studies and a critical assumption in the use of regression models. We trust that the preceding discussion will enable readers to identify the possibility for dependence in a study design, and thereby furnish a starting point for choosing a suitable method of analysis.

14.4. Multi-State Models

14.4.1 The Study

Another example of a study that resulted in longitudinal data is an investigation of disease progression in patients with psoriatic arthritis reported by Gladman et al. [30]. The subjects enrolled in the study were followed prospectively over a period of 14 years. During this time, study participants were treated at a single clinic and, at each clinic visit, standardized assessments of clinical and laboratory variables were obtained.

To characterize the progression of a patient's disease during the study, investigators chose to measure the number of joints that were permanently damaged. Since this outcome variable was evaluated at each clinic visit, dependence between observations on the same study participant is once again a prominent feature of the design of this study. Although a Liang-Zeger regression model could be used to analyze these data, sometimes it is desirable to model temporal changes in the outcome variable more directly than the simple inclusion of a correlation parameter affords. In this case, for example, the number of damaged joints can only increase in time, and an appropriate statistical model should reflect this fact.

The study investigators measured a number of different explanatory variables whose relationship with disease progression could be examined. In the following discussion, we will focus on two of these measures. The first, X_1, indicates the number of effusions, i.e., inflamed joints, while the second, X_2, is derived from the laboratory measurement of erythrocyte sedimentation rate (ESR). The binary variable, X_1, is equal to one if the number of effusions exceeds four, and is coded zero otherwise. Likewise, if the ESR rate is less than 15 mm/h then the binary covariate, X_2, is equal to one; otherwise, $X_2 = 0$. The values of these covariates were determined for each study participant at the time of the first clinic visit since the primary goal in this study was to determine if patients at an increased risk of disease progression could be identified early in the course of their disease. The identification of such patients would facilitate earlier treatment intervention and also might allow physicians to avoid the unnecessary use of aggressive treatment in patients expected to experience a more benign course of the disease. In other study settings, interest might focus on the relationship between outcome and explanatory variates where both types of observations are measured repeatedly over time. In principle, the methods that we describe in the remainder of this section can be extended to include this more complicated type of analysis.

14.4.2. A Multi-State Statistical Model

The method of analyzing longitudinal data that we intend to use for this study of psoriatic arthritis patients is called multi-state modelling. This approach is possible when, at any point in time, each study subject can be classified into exactly one of a set of possible states. For example, carefully defined conditions such as 'well', 'ill', and 'dead' could represent the set of possible states of interest in a longitudinal study. In other investigations, including this study of psoriatic arthritis, the states may represent different values of an outcome variable of interest. In this study, the states in the model are determined by the number of damaged joints. The state labelled 1 denotes no damaged

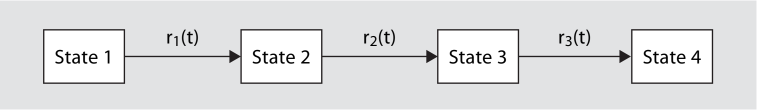


Fig. 14.1. Schematic representation of the multi-state model used to analyze the psoriatic arthritis data.

joints; similarly, States 2, 3, and 4 correspond to one to four, five to nine, and ten or more damaged joints, respectively.

The analysis of multi-state models is based on movements, that we will call transitions, from one state to another. Each transition is modelled, statistically, in a manner similar to that used in chapter 13 to analyze survival data. Therefore, a multi-state model involves a rate for each possible transition between states. Since joint damage is a progressive condition, the only possible transitions are from State 1 to State 2, State 2 to State 3, and State 3 to State 4. We will denote a transition rate at time t by the function $r_i(t)$; the subscript i indicates that $r_i(t)$ is the transition rate from the state labelled i to the state labelled i+1. Therefore, the multi-state model for the psoriatic arthritis study involves the three rate functions $r_1(t)$, $r_2(t)$,and $r_3(t)$.

Figure 14.1 summarizes the structure of this multi-state model of the study in a schematic way. The boxes represent the different states and arrows indicate the various transitions that can occur; the corresponding transition rates are indicated above the arrows.

In the model for a hazard rate used in chapter 13 we incorporated the possible effects of various covariates. To achieve the same goal in this multi-state model for the psoriatic arthritis study, we represent the transition rate between State i and State i+1 for a subject with observed values of the covariates $X_1 = x_1, \ldots, X_k = x_k$, or $\underset{\sim}{x} = \{x_1, x_2, ..., x_k\}$, by the equation

$$\log\{r_i(t; \underset{\sim}{x})\} = \log(a_i) + b_{i1}x_1 + b_{i2}x_2 + \ \dots \ + b_{ik}x_k.$$

In contrast to the hazard rate models that we used in chapter 13, this transition rate model has more than one regression coefficient associated with each covariate. In fact, for each covariate, X_j say, there is a different regression coefficient for each transition rate function in the multi-state model. This is why each regression coefficient, b_{ij}, has two subscripts. The first subscript, i, identifies the state from which the transition occurs and the second subscript, j, denotes the corresponding covariate, X_j. Although the number of regression coefficients associated with each covariate has increased, the interpretation of regression coefficients is just as we have previously described for all other regression models. A non-zero value for b_{ij} indicates that the covariate, X_j, is associated with the occurrence of transitions from State i to State i+1. For the transition from State i to State i+1, the relative risk, or ratio of transition rates, for an individual with $X_j = 1$ compared to an individual with $X_j = 0$ is $\exp(b_{ij})$.

This multi-state model does not incorporate the correlations between observations on the same subject directly. However, the correlation has little impact in this situation because the analysis is based on non-overlapping portions of the follow-up period. To estimate the regression coefficients associated with transitions from State 2 to State 3, say, only the information pertaining to subjects who occupy State 2 is used. What happens to a subject before or after that individual occupies State 2 does not contribute to the estimation of the regression coefficients $b_{21}, b_{22}, ..., b_{2k}$.

Notice also that the term a_i does not depend on the length of time a study subject has occupied State i. This assumption is not necessary, but it simplifies the analysis considerably; moreover, it is frequently possible to define states in such a way that this simplifying assumption is a reasonable approximation.

The technical name for data that are used to estimate the regression coefficients in this multi-state model is panel data. Individuals are observed on different occasions and, at each of these times, the state that each subject occupies is recorded. The requirements for panel data do not dictate that the observation times must be equally spaced. The information that is used to estimate the regression coefficients in the model for the transition rate, $r_i(t)$, is whether, between any two consecutive observation times, an individual has moved from State i to another state or remained in State i. If a transition has occurred, the estimation procedure takes into account all possible intermediate states through which the subject might have passed. Additional details concerning the analysis of panel data may be found in Kalbfleisch and Lawless [31].

If it is reasonable to assume that the regression coefficients associated with a single covariate are all the same, the analysis of this multi-state model can be simplified considerably. In the psoriatic arthritis study, this assumption would correspond to specifying that $b_{1j} = b_{2j} = b_{3j}$ for each covariate, X_j, $j = 1, 2, ..., k$. However, this assumption must be tested and an appropriate significance test can be devised for this purpose. There is no standard terminology for such a test; we will refer to it as a test for common covariate effects. The corresponding assumption of common covariate effects is frequently made in the early stages of an investigation when the effects of many covariates are being investigated; the primary aim of such a procedure is to screen for covariates that warrant further attention. At the latter stages of an analysis, the test for common covariate effects is used to simplify models that involve multiple covariates.

Table 14.3. Observed changes in damage states between consecutive clinic visits for the psoriatic arthritis study

Transition	Entry state		
	1	2	3
1 → 2	56		
1 → 3	16		
1 → 4	8		
2 → 3	12	23	
2 → 4	3	15	
3 → 4	13	11	13
Total number of patients	213	75	17

14.4.3. Illustrative Results

When the data from the psoriatic arthritis study were compiled for analysis, 305 patients who had fewer than ten damaged joints when they became patients were registered at the clinic. As we indicated in the previous subsection, estimation of the regression coefficients in the multi-state model is based on the number of transitions that occurred between the various states in the model during the follow-up period for each subject. In this study, patients entered the clinic in States 1, 2, and 3. Some patients did not experience a change in state after becoming clinic patients; other individuals experienced more than one change of state. Table 14.3 summarizes the various transitions that occurred between consecutive visits to the clinic.

Because subjects occupied different states when they became clinic patients and entered the study, it would be desirable to incorporate the possibility of differences between patients who enter the study at different stages in the disease course. The model that we outlined in the previous section can be generalized to accommodate this possibility by stratifying the analysis according to the entry state for each subject. The resulting stratified multi-state model is analogous to the stratified regression models for the hazard function that we previously discussed in chapter 13. To stratify our analysis of the psoriatic arthritis data according to the states occupied by subjects when they became clinic patients, we simply need to allow the a_i terms in the regression model for $r_i(t; \underset{\sim}{x})$ to depend on the entry state. We will continue to assume that the effects of the various covariates on the rate of transitions from State i are common across the strata. Hereafter, we will focus our attention solely on the regression coefficients, i.e., the b_{ij}s. However, in each analysis that we describe, the stra-

Table 14.4. The results of a stratified, multi-state regression analysis of disease progression in psoriatic arthritis patients

Covariate	Transition	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
Effusions	1 → 2	0.84	0.34	2.47	0.013
	2 → 3	0.55	0.22	2.50	0.012
	3 → 4	0.42	0.25	1.68	0.093
ESR	1 → 2	−0.31	0.23	1.35	0.177
	2 → 3	−0.71	0.23	3.09	0.002
	3 → 4	0.39	0.31	1.26	0.208

tum-specific effects that are represented by three a_i terms, one for each possible entry state in the study, also have been estimated.

Table 14.4 summarizes the estimated regression coefficients and corresponding standard errors for a stratified multi-state model that incorporates the two explanatory variables X_1 and X_2. Recall that X_1 is a binary covariate indicating that the subject had more than four joints with effusions at the first clinic visit, and X_2 is another binary covariate that indicates the subject had an ESR that was less than 15 mm/h. Three regression coefficients were estimated for each covariate; these coefficients correspond to the effects of the covariate on transitions from States 1, 2, and 3.

According to the results presented in table 14.4, both X_1 and X_2 are associated with one or more transition rates. The estimated regression coefficients for X_1, the number of effusions, are roughly comparable. For X_2, the covariate indicating ESR at the first clinic visit, the estimated regression coefficients corresponding to transitions from State 1 to State 2 and State 2 to State 3 are comparable whereas the estimated regression coefficient for transitions from State 3 to State 4 has a different sign and is not significantly different from zero.

The results of this preliminary analysis of the psoriatic arthritis data suggest that a simpler model may fit the observed data as well as the preliminary model summarized in table 14.4. This simpler model incorporates a common effect for X_1 across all transitions and a common effect for X_2 for transitions from States 1 and 2 only. When this simpler model is used to analyze the psoriatic arthritis data, the estimated regression coefficient for X_1 is 0.53 with a corresponding standard error of 0.14. The estimated regression coefficient associated with X_2 is −0.52; its estimated standard error is 0.16. Although it is beyond the scope of this book, a significance test can be carried out to evaluate

Table 14.5. Estimated relative risks for the simpler version of the multi-state regression model analyzed in table 14.4

Covariate	Transition		
	1 → 2	2 → 3	3 → 4
Effusions	1.70	1.70	1.70
ESR	0.59	0.59	

whether, when the two models are compared, the simpler model provides an adequate summary of the data. This test leads to an observed value for the test statistic that is equal to 2.8. If the null hypothesis concerning the adequacy of the simpler model is true, this test statistic should follow a χ_4^2 distribution. Therefore, the significance level of the aforementioned test is approximately 0.60, and we conclude that the simpler multi-state model adequately represents the psoriatic arthritis data.

To summarize the results of fitting this multi-state model, it may be useful to tabulate the estimated relative risks corresponding to the effects of the explanatory variables; see table 14.5. Often, relative risks can be interpreted more readily than the corresponding estimated regression coefficients.

Based on the results of this simplified analysis of the psoriatic arthritis data, the findings of this longitudinal study would be that evidence of extensive inflammation at the first clinic visit is associated with more rapid progression of joint damage in the future. However, a low sedimentation rate when the patient is first evaluated is related to a lower rate of disease progression subsequently, provided the number of damaged joints at the first visit is small.

14.4.4. Conclusions

The use of multi-state models to analyze longitudinal medical studies is increasing. A number of features of the model we have used in this section go beyond what we can describe adequately in this book. However, we hope that our description of the psoriatic arthritis study and its analysis has provided some indication of how multi-state models are related to simpler regression models and of their potential value in the statistical analysis of longitudinal data.

In the last five chapters, we have presented a brief introduction to the concept of regression models. A much longer exposition of these topics would be required to enable readers to fully master these important techniques. Nevertheless, we trust that our discussion has helped to provide some understanding of these methods, and has rendered their use in research papers somewhat less mysterious.

15

Analysis of Variance

15.1. Introduction

In §10.3 we noted that the results of fitting linear regression models can be summarized in analysis of variance (ANOVA) tables, and §10.5 provided a brief introduction to such tables. The brevity of that explanation was largely because ANOVA tables are not particularly relevant to the other types of regression models that we discussed in chapters 11 through 14. However, in spite of its link to linear regression models, ANOVA is sometimes treated as a specialized statistical technique, especially when it is used to analyze carefully designed experiments. Therefore, we now return to this topic to provide a more extended explanation that it rightly deserves.

The use of ANOVA methods is often linked to the special case of multiple linear regression in which the explanatory variables are categorical measurements, rather than continuous, or approximately continuous. The body weights of mouse pups and the sizes of the litters in which those pups were nurtured that we used in the example discussed in chapter 10 are typical examples of such quantitative measurements. However, provided we devote some care to how we choose to represent categorical information like type of treatment, sex, tumor grade, etc., we can include these and other details about subjects that are recorded during a study as explanatory variables in multiple regression models.

Table 15.1 summarizes the results of an ANOVA that was prepared to investigate the simultaneous effect of factors identified by the labels Machine and Reagent on the logarithmic international normalized ratio (log INR) value observed in 354 specimens. The study in question was an investigation of the possible effect of three different types of laboratory measurement equipment, and two distinct thromboplastin reagents, on the prothrombin time measure-

Table 15.1. ANOVA table corresponding to a regression of the logarithmic international normalized ratio (INR) from 354 specimens on explanatory variables representing machine and reagent study conditions

Term	SS	DF	MS	F	Significance level
Machine	1.581	2	0.790	16.1	<0.001
Reagent	2.598	1	2.598	52.9	<0.001
Machine × Reagent	0.087	2	0.044	0.88	0.42
Residual	17.107	348	0.049		
Total	21.373				

ment for a diverse spectrum of patients, using the INR methodology first adopted by the WHO in 1973.

Using only the insight garnered from our previous discussion of such summary tables in §10.5, we might infer that two of the three terms in the fitted regression model, i.e., the ones called Machine and Reagent, constitute important, systematic sources of variation in the log INR measurements obtained from the 354 specimens tested. However, these are clearly categorical variables, involving obvious subjectivity in how they might be labelled, rather than quantitative measurements like the body weight of a mouse pup. To be able to understand table 15.1 and how it relates to a corresponding regression model better, we need to know how information such as the lab machine used to obtain an INR measurement can be suitably represented in a regression model equation.

The methodology known as ANOVA can be used to analyze very complex experimental studies; indeed, entire books have been written on the subject. Our goal in this chapter will be a very modest one. By building on the understanding of multiple regression that readers have already acquired, we aim to furnish a wider perspective on this oft-used set of tools which could form the basis for a more general appreciation of its use and for additional reading on the topic if desired.

15.2. Representing Categorical Information in Regression Models

Although we have not previously used categorical information like Machine or Reagent details in any of the regression models that we considered in chapter 10, a brief glance at figure 15.1 will convince most readers that the INR

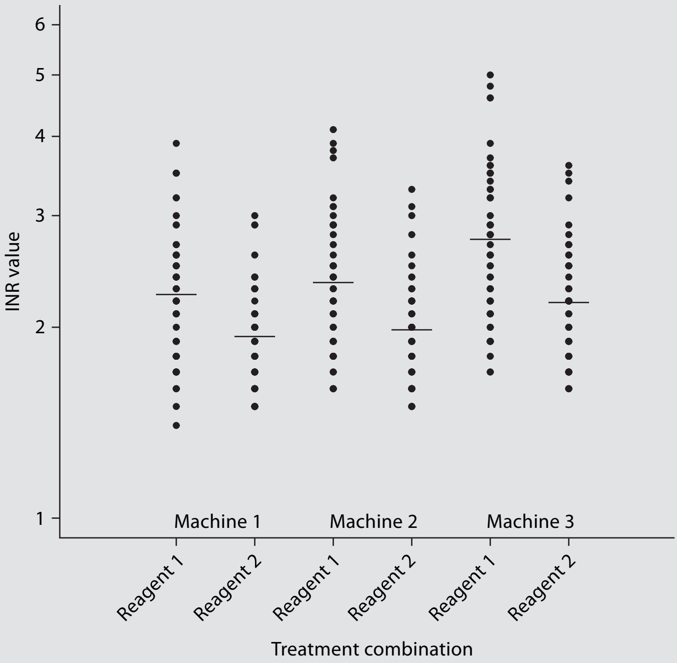


Fig. 15.1. Dependence of the prothrombin time international normalized ratio (INR) measurements on the six experimental conditions determined by the machine and thromboplastin reagent used. The horizontal lines indicate the average INR measurement for that combination.

measurement for any particular specimen surely depends on aspects of the measurement process, e.g., the choices of Machine and Reagent used, in some systematic way. If we can identify and estimate these effects, we would begin to understand the measurement system that produces INR values, and perhaps gain some insights concerning how that process might be maintained, or even improved. We could adopt the representations used in figure 15.1 and label the machines 1, 2 and 3, and the reagents 1 and 2, but these assigned numbers are purely convenient labels. With no less justification, we could in-

Table 15.2. Settings of the indicator variables, X_1, X_2, and X_3, that identify machine and reagent information in the INR study

X_1	X_2	X_3	Reagent	Machine
1	0	0	1	1
1	1	0	1	2
1	0	1	1	3
0	0	0	2	1
0	1	0	2	2
0	0	1	2	3

stead label the machines A, B and C, and the reagents M and N. If labels for categories are so arbitrary, how can we incorporate categorical information into a multiple regression model as one or more explanatory variables, since the values of all the variables used in the multiple regression calculations must be numbers?

One way to achieve our goal involves using sets of appropriate indicator variables, i.e., variables that equal either 0 or 1 according to well-defined rules. For example, a single indicator variable would suffice to identify whether an INR measurement was obtained using the reagent labelled 1. For convenience, let's call that indicator variable X_1. So if the value of X_1 associated with a particular INR measurement is 1, we know immediately that reagent 1 was used to obtain that INR measurement. And if the value assigned to X_1 happens to be 0, then we know that reagent 2 was used instead. Hence the values that we assign to X_1 for each INR measurement will identify which reagent was used to obtain each of the 354 observations in the study.

To identify the three separate machines used in the investigation we require two additional indicator variables; let's call them X_2 and X_3. Table 15.2 shows how these two variables, together with X_1, could unequivocally encode the machine and reagent information associated with any particular INR measurement.

Table 15.2 implicitly defines how X_2 and X_3 encode the lab equipment details. The combination $X_2 = 0$, $X_3 = 0$ identifies machine 1, $X_2 = 1$, $X_3 = 0$ corresponds to machine 2, and machine 3 is represented by the pair of values $X_2 = 0$, $X_3 = 1$.

It is worth noting that the encoding summarized in table 15.2 is not unique. For example, we could switch assignment of the values 0 and 1 for X_1

Table 15.3. Estimated regression coefficients and standard errors for a multiple linear regression of logarithmic INR on the reagent and machine indicator variables X_1, X_2 and X_3

Explanatory variable	Estimated regression coefficient	Estimated standard error
X_1	0.171	0.024
X_2	0.032	0.029
X_3	0.155	0.029

Table 15.4. ANOVA table corresponding to the regression of logarithmic INR on the reagent and machine indicator variables X_1, X_2 and X_3

Term	SS	DF	MS	F	Significance level
Machine	1.581	2	0.790	16.1	<0.001
Reagent	2.598	1	2.598	52.9	<0.001
Residual	17.194	350	0.049		
Total	21.373				

and the two reagents it identifies. We could also link the assignments of X_2 and X_3 to the three machines in a different way than table 15.2 specifies. However, if we are careful in defining our sets of indicator variables, we can always produce a representation for categorical information that will be satisfactory, and that can be reinterpreted in terms of the original category labels. Moreover, since indicator variables have numerical values – either 0 or 1 – we can multiply them by suitable regression coefficients, such as b_1, b_2 and b_3, and incorporate such products into a multiple regression model equation, such as one for logarithmic INR in our example. Thus, we might want to use the regression model that corresponds to the equation

$$\log \mathrm{INR} = Y = a + b_1X_1 + b_2X_2 + b_3X_3,$$

in which case the estimated regression coefficients, and the corresponding estimated standard errors, would be those summarized in table 15.3.

The corresponding ANOVA is displayed in table 15.4. As we explained in §10.5, the sum of squares (SS) associated with the line labelled Total can always be divided into two parts. In §10.5, this partition was represented as

$$SS_{Total} = SS_{Residual} + SS_{Model}.$$

The first sum of squares on the right-hand side of this equation obviously corresponds to the relevant entry in the line of the ANOVA table labelled Residual. The second, SS_{Model}, represents the sum of squares for the remaining lines of the ANOVA table. The residual sum of squares, i.e., $SS_{Residual}$, represents variation in log INR values that is not explained by the regression model; SS_{Model} is the variation in log INR values that is explained or accounted for by the model, that is, by the machine and reagent information encoded in the values of X_1, X_2, and X_3. The value of SS_{Model} can be subdivided in various ways to relate to specific explanatory variables. The details of the various ways to do this are beyond the scope of this book but, as in §10.5, we can discuss such a refined subdivision in general terms for this particular example.

The combined contributions associated with the variables X_2 and X_3 are amalgamated in the line labelled Machine, since these two indicator variables jointly encode the details about which machine was used to obtain each INR measurement. The DF column entry in table 15.4 for Machine is equal to 2 because each of the two associated indicator variables has a corresponding estimated regression coefficient. The contribution associated with the variable X_1 is represented in the line labelled Reagent and has only 1 as its corresponding DF entry since the only associated regression coefficient is b_1.

It is worth noting that, in common with tests associated with other regression models, the significance tests recorded in table 15.4 relate to the relationship between Machine and log INR after adjusting for Reagent and between Reagent and log INR after adjusting for Machine. This is true, however, only because the experiment is a carefully designed one with appropriate balance in the number of observations for each Machine and Reagent combination. Section 10.5 discusses the more general approach to such tests in ANOVA tables derived from less controlled data collection. Recall also that a significance test concerning the importance of a single explanatory variable, such as X_1, in a regression model is comparable to the usual test for a non-zero regression coefficient in any regression model, based on results such as those for X_1 found in table 15.3.

In terms of the regression coefficients b_1, b_2 and b_3, the F-test associated with Reagent is therefore a test of the null hypothesis $b_1 = 0$. The F-test associated with Machine is a test of the null hypothesis that both b_2 and b_3 are zero, i.e., $b_2 = b_3 = 0$. The latter test is an example of what is sometimes called a 'global' test of the relationship between a response or outcome variable and a cate-

Table 15.5. Estimated mean logarithmic INR values for the six combinations of reagent and machine used in the INR measurement study, based on the fitted regression model $\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3$

Machine 1		Machine 2		Machine 3	
Reagent 1	Reagent 2	Reagent 1	Reagent 2	Reagent 1	Reagent 2
0.799	0.628	0.831	0.660	0.954	0.783

gorization of observations into more than two groups or classes where no natural ordering of those groups is assumed.

From the estimated regression coefficients summarized in table 15.3, and the estimated value of a, which is 0.628, we could calculate the mean logarithmic INR measurement for each of the six combinations of lab machine and reagent used in the study. For example, $X_1 = 0$, $X_2 = 0$, $X_3 = 0$ identifies the use of machine 1 and reagent 2 and, according to the fitted regression model, the estimated mean value of log INR is $\hat{Y} = \hat{a} + \hat{b}_1 (0) + \hat{b}_2 (0) + \hat{b}_3 (0) = 0.628$. The complete set of estimated mean values is summarized in table 15.5. Readers will recognize that the ranking and rough magnitude of these estimated means correspond to the approximate positions of the six horizontal lines in figure 15.1 that mark the average INR values for the various equipment combinations.

The regression model involving X_1, X_2 and X_3 is known as a main effects model for logarithmic INR. The indicator variable X_1 represents the main effect of thromboplastin reagent on the mean value of the logarithmic INR. Changing the value of X_1 from 0 to 1 corresponds to the change predicted by the model in the mean value of logarithmic INR that results when we replace reagent 2 by reagent 1. Similarly, the pair of indicator variables X_2 and X_3 jointly represent the main effect on the mean value of logarithmic INR, predicted by our regression model, of changing the lab machine used in the measurement process. Since the values $X_2 = 0$ and $X_3 = 0$ jointly designate machine 1, changing the value of X_2 from 0 to 1, while holding X_3 fixed at 0, corresponds to the change in the mean value of logarithmic INR when Machine 1 is replaced by Machine 2 in the INR measurement process. Or holding the value of X_2 fixed at 0 but changing X_3 from 0 to 1 is equivalent to replacing Machine 1 by Machine 3 in the measurement system. As a rule, if a categorical explanatory variable involves k distinct category types – which are usually referred to as levels – the main effect of that variable can be represented in a regression model by a set of k – 1 suitably defined indicator

Table 15.6. Settings of the indicator variables, S1, S2, S3, S4, and S5, that could also encode machine and reagent information in the INR study

S1	S2	S3	S4	S5	Reagent	Machine
0	0	0	0	0	1	1
1	0	0	0	0	2	1
0	1	0	0	0	1	2
0	0	1	0	0	2	2
0	0	0	1	0	1	3
0	0	0	0	1	2	3

variables. Therefore, since Reagent has two levels, namely 1 and 2, and Machine has three levels, i.e., 1, 2 and 3, we have represented the main effect called Reagent by X_1 and the corresponding main effect labelled Machine by X_2 and X_3.

Before we proceed to introduce the concept known as two-factor interactions, it is worth mentioning that the indicator variables we have used to represent categorical information are not the only encoding method we could have chosen. Indeed, there are several well-known representations for a factor with k levels, including one known as orthogonal contrasts. Describing these alternatives is a pedagogical challenge that we intend to forego; undoubtedly, it is well beyond the scope of this book. We only mention this detail because readers may encounter it elsewhere.

15.3. Understanding Two-Factor Interactions

Since the INR study involved six distinct treatment combinations, it is possible, and some readers might prefer, to define a single categorical variable with six levels that correspond to the six lab machine and reagent combinations. In that case, we could have chosen to use five indicator variables – let's call them S_1, S_2, S_3, S_4, and S_5 – to represent the six treatment combinations of lab machine and reagent. (We use the letter S to denote the single categorical variable, and to differentiate the associated five indicator variables from the previous three labelled X_1, X_2 and X_3.) Table 15.6 summarizes an encoding of machine and reagent information using the values of the five subscripted S variables.

Table 15.7. Estimated mean logarithmic INR values, based on the linear regression model that uses the surrogate indicator variables S1 through S5 to encode machine and reagent information

Machine 1		Machine 2		Machine 3	
Reagent 1	Reagent 2	Reagent 1	Reagent 2	Reagent 1	Reagent 2
0.783	0.644	0.826	0.665	0.975	0.762

If we chose to use these five indicator variables, and the associated representation of machine and reagent information in a multiple regression of the logarithmic INR measurements on S_1, S_2, S_3, S_4, and S_5, the model equation would involve five coefficients and correspond to the formula

$$\log \text{INR} = Y = a + b_1S_1 + b_2S_2 + b_3S_3 + b_4S_4 + b_5S_5.$$

Although we won't bother to summarize the estimated regression coefficients and standard errors for this alternative model, table 15.7 indicates the estimated mean values, $\hat{Y}$, of log INR measurements predicted for the six combinations of machine and reagent.

Like the estimates displayed in table 15.5, these predicted mean logarithmic INR values also approximate the relative position and magnitude (on the logarithmic scale) of the horizontal lines displayed in figure 15.1. However, readers will note that these two competing regression models fitted to the logarithmic INR measurements do not give the same predictions and therefore are not the same. The so-called main effects model, based on the three indicator variables we called X_1, X_2 and X_3, results in three estimated regression coefficients and the predicted mean logarithmic INR values displayed in table 15.5, whereas the alternative regression model based on the five indicator variables we called S_1, S_2, S_3, S_4, and S_5 produced five corresponding estimated regression coefficients and the predicted mean logarithmic INR values summarized in table 15.7.

The main effects regression model, which uses the variables X_1, X_2 and X_3, represents a simplification of the alternative regression model, based on S_1 through S_5, and leads to predicted mean values that satisfy certain predictable mathematical restrictions. For example, regardless of the machine involved, the difference between the estimated mean log INR using reagent 1 and the corresponding mean log INR using reagent 2 is always $0.171 = 0.799 - 0.628 = 0.831 - 0.660 = 0.954 - 0.783$ log units. The same restrictions don't apply to the alternative regression model because we chose to use a larger set – to be precise, five – of indicator variables to encode the reagent and machine information. As a result, the regression model involving S_1 through S_5 is more flexible, but that flexibility comes with a price. The model is also a more complex characterization of the INR measurement study, and requires us to estimate five regression coefficients rather than three. In fact, this regression model includes the Machine × Reagent two-factor interaction which is the third row of entries found in table 15.1. This two-factor interaction is, however, more conventionally represented mathematically by the two additional indicator variables $X_4 = X_1 \times X_2$ and $X_5 = X_1 \times X_3$. If we had fitted a multiple regression model to the logarithmic INR measurements that involved X_1, X_2, X_3, X_4, and X_5, the resulting table of predicted mean log INR values would have been identical to table 15.7.

Which model is preferable for the INR data? The answer to that question can be found in the row of table 15.1 labelled Machine × Reagent. The significance test associated with this line of the table is a test of the need to include the associated explanatory variables in the model or, equivalently, a test of the null hypothesis that the regression coefficients b_4 (associated with X_4) and b_5 (corresponding to X_5) are both equal to 0. Since this test does not have a small significance level, we would be adopting a characterization of the INR measurement study that was unnecessarily complex if we chose to use the larger set of indicator variables, X_1, X_2, X_3, X_4, and X_5, that includes the Machine × Reagent two-factor interaction, or the equivalent model based on S_1, S_2, S_3, S_4, and S_5, rather than simply the two main effects represented by the three indicator variables, X_1, X_2 and X_3, to encode the machine and reagent information. For these data, it appears that the mathematical restrictions associated with a purely main effects model will not prevent us from appropriately characterizing the impact, on mean logarithmic INR, of changing either the thromboplastin reagent or the laboratory machines used to determine INR measurements on the logarithmic scale.

15.4. Revisiting the INR Study

Thus far in our investigation of the INR study, we have intentionally overlooked an essential feature of the study design, so that we could focus attention on the distinction between a main effects model and one involving an interaction between the two main effects known as Machine and Reagent.

The 354 distinct INR measurements that we analyzed actually came from 59 patients whose prothrombin times varied quite markedly. Every patient's plasma sample was subdivided into six specimens, and each of these specimens was then measured using one of the six machine/reagent combinations that

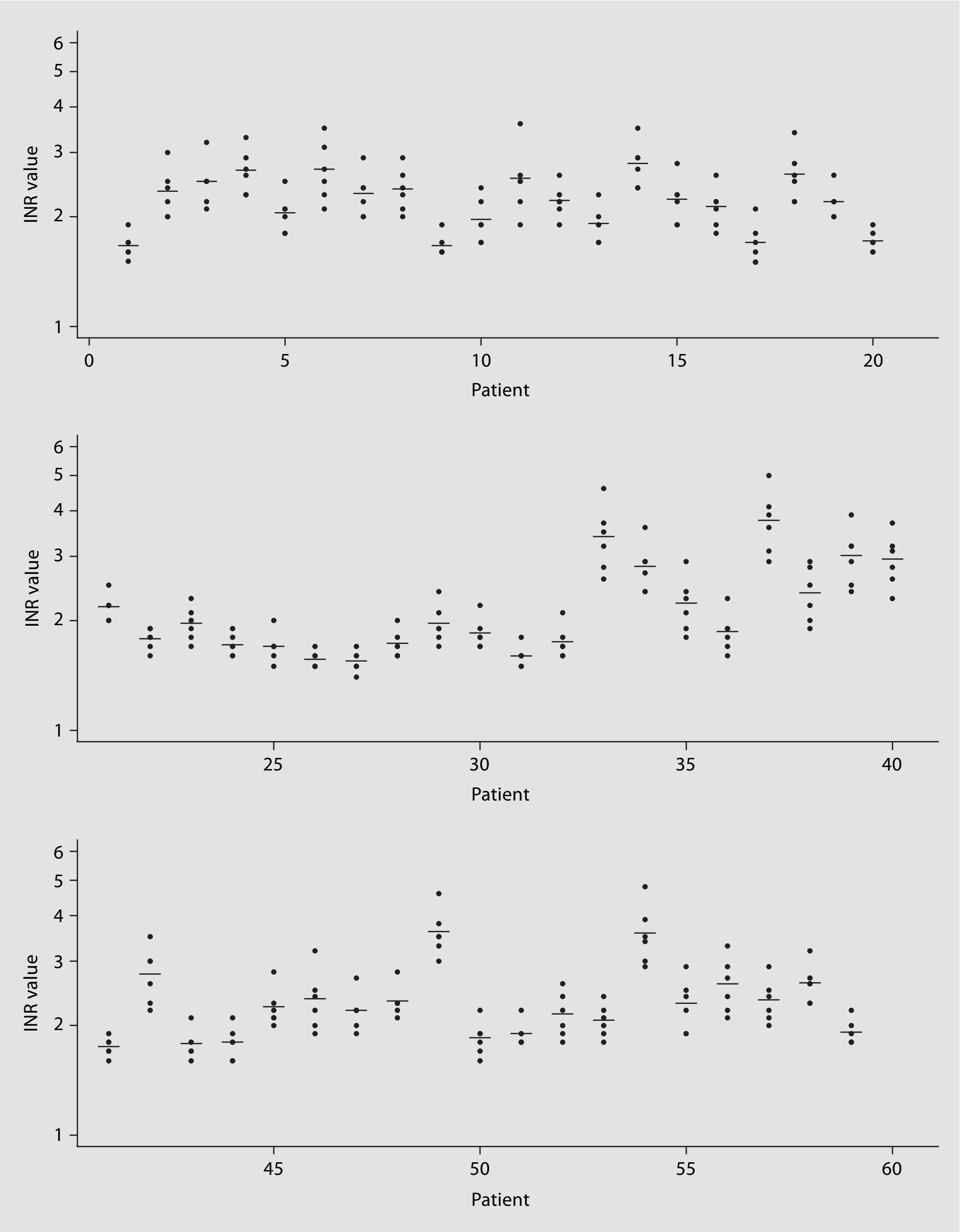


Fig. 15.2. Repeated measurements obtained from 59 subjects during an experimental study of the prothrombin time INR measurement system. The horizontal lines indicate patient-specific average values.

was a focus of the study. Thus, the investigators had 6 × 59 = 354 units with which to work. Figure 15.2 illustrates graphically how the resulting prothrombin time measurements varied. Obviously, the investigators suspected that patient-to-patient variation in the measured INR values might drown out any subtle clues concerning the roles played by lab equipment and thromboplastin reagent in the INR measurement process unless they designed the study carefully. By subdividing each patient's specimen into six identical units, and measuring the associated INR value using each of the six possible combinations of machine and reagent, the researchers created the six-way analog of a study with natural pairing in the sense that we have six observations per patient rather than the two we would have, for example, in data used for a paired t-test, such as the one discussed in §9.4.1. Among the six measurements of INR obtained from a single patient's subdivided serum, only changes in the lab equipment and reagent used are plausible explanations for systematic variation in the corresponding INR values. Thus, we should in fact make comparisons between Machine and Reagent combinations only by comparing different measurements from the same patient. The results from these comparisons for each patient are then combined. This is known as making comparisons 'Within Patients'.

There is a parallel between the structure of this carefully designed experiment and that which we encountered in chapter 14 during our consideration of data derived from longitudinal studies. As we mentioned there, a key concern is the need for any analysis to adjust for or take into account appropriately the correlation between repeat observations from the same study subject. In chapter 14 we focussed on repeat observations over time, but repeat observations can arise in other situations. The method that we will outline here, in the context of a linear regression model, is a special case of what is known as *repeated measures analysis of variance.* The basic approach derives from the introduction of a 'Patient' effect into the regression model. Thus far the two regression models we have used to analyze the INR data have not included any terms representing patient-to-patient variability.

To represent the categorical information that corresponds to some convenient labelling of the 59 patients, we can define 58 suitable indicator variables, and include all 58 in the regression model as well. Suppose we call those 58 indicator variables $P_1, P_2, \ldots, P_{58}$, for convenience. We should definitely include these 58 indicator variables representing patient labels in the regression model because of the particular way in which the INR measurements were collected. The choice not to do so would be equivalent to ignoring an essential feature of the experiment, the correlation between observations from the same patient. There is no need, and in fact it is not possible, to test whether the regression coefficients corresponding to these 58 variables are all equal to zero. These

Table 15.8. ANOVA table corresponding to the regression of the logarithmic international normalized ratio (INR) on explanatory variables representing patient, machine and reagent study conditions

Term	SS	DF	MS	F	Significance level
Between Patients	16.230	58	0.280		
Within Patients					
Machine	1.581	2	0.790	239.6	<0.001
Reagent	2.598	1	2.598	787.5	<0.001
Residual	0.963	292	0.003		
Total	21.373				

coefficients are linked to the mean log INR values for the six measurements from each patient, and variation in these means corresponds to 'Between Patients' variability.

Table 15.8 is the ANOVA summary that results from including these additional 58 indicator variables in the linear regression model. The structure of this ANOVA table is based on the relationship

$$SS_{\text{Total}} = SS_{\text{Between Patients}} + SS_{\text{Within Patients}}.$$

In this revised ANOVA, patient-to-patient variability in the mean log INR values for the 59 patients is represented by the value of the Between Patients entry in the column labelled SS. This has been separated from the Residual line in which it was previously hidden when the associated regression model did not include a set of indicator variables representing the patient label information (compare tables 15.4 and 15.8). The entry labelled Between Patients in the sum of squares column of the table is based on the squared differences between individual patient means and the overall sample mean $\bar{Y}$; the corresponding entry in the column labelled MS can be regarded as an estimate of the variance of these mean values. When patients can be classified in one or more ways, e.g., as Treatment and Control groups, perhaps, it is possible to extract, from the Between Patients SS, values for the various sums of squares that relate to such classifications. The remainder of the Between Patient SS can then be used to define another type of Residual SS for Between Patient comparisons that is based on these patient means, but we do not propose to pursue this more complicated possibility here.

The SS entries in table 15.8 under the heading Within Patients represent a subdivision of the second component of the Total SS, the one that we previously referred to as $SS_{\text{Within Patients}}$. Notice that the sums of squares labelled Machine and Reagent are still equal to 1.581 and 2.598, respectively. These are the same numerical values that we first encountered in table 15.1. The fact that their values are identical in both ANOVA tables is due to the underlying balance in the study design. However, by including a main effect for Patient, i.e., P_1, P_2, …, P_{58}, in our revised regression model, we have decreased the mean square in the Residual line from 0.049 to 0.003, which represents roughly a 15-fold reduction. However, this Within Patients residual line in the ANOVA table is now only a suitable basis for significance tests involving comparisons that are based on within-patient data. Thus, table 15.8 now essentially contains an ANOVA sub-table that is relevant only to within-patient comparisons.

As a result of these changes, the F-ratios for Machine and Reagent listed in table 15.8 are noticeably larger than the corresponding entries in table 15.4. Although any conclusions concerning the importance of the individual machine and reagent effects on the INR measurement process would be roughly similar if we were to rely on the analysis summarized in table 15.4, readers should see immediately that by including patient information in the regression model and thereby reducing the Residual SS, we have greatly enhanced our ability to identify non-zero main effects associated with the primary factors in the study design. Moreover, this enhanced statistical ability to distinguish between signal and noise should prompt us to re-examine the possible importance of the Machine × Reagent two-factor interaction. The sum of squares associated with including the corresponding indicator variables, $X_4 = X_1 \times X_2$ and $X_5 = X_1 \times X_3$, in the regression model was previously identified as 0.087. If this value is divided by 2, its degrees of freedom, to derive the associated mean square, the resulting F-ratio could easily be a tenfold multiple of the now-reduced residual mean square value of 0.003.

The revised ANOVA summary – see table 15.8 – for the regression model that includes a main effect for Patient (represented by the indicator variables P_1, P_2, …, P_{58}), as well as main effects for Machine and Reagent, and the two-factor interaction between machine and reagent (represented jointly by the five indicator variables X_1, X_2, X_3, X_4, and X_5), reveals that the interaction is now associated with a very small significance level. Therefore, the experimental data constitute evidence that the mean values of logarithmic INR depend in a complex way on the combination of lab machine and thromboplastin reagent used to measure patient clotting time.

In our previous discussion of two-factor interactions, we indicated that a simple model which involves only the main effects for Machine and Reagent, in addition now to the very essential main effect that represents patient-to-patient variability, will impose certain mathematical restrictions on the estimated mean values of the corresponding logarithmic INR measurements. Inclu-

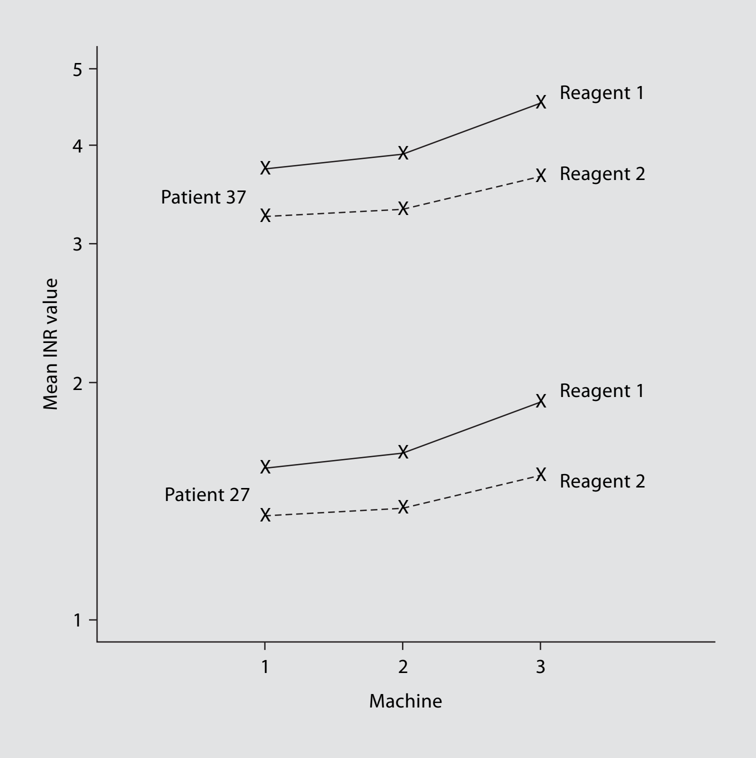


Fig. 15.3. Predicted mean INR values for patients 27 and 37, based on the ANOVA summarized in table 15.9.

sion of the Machine × Reagent two-factor interaction relaxes those mathematical restrictions, allowing the six estimated means identified with the six machine/reagent combinations to conform to the dictates of the study data.

Figure 15.3 displays estimated mean INR values obtained from the fitted regression model for each machine and reagent combination. The two sets of values plotted on the graph correspond to those for patients 27 and 37, whose predicted mean INR values represented the extremes occurring in the study. The solid and dashed lines have been added only to enhance visual appreciation; obviously, values for Machine between the labels 1, 2 and 3 have absolutely no sensible meaning. From the estimated means denoted by the character 'X' on the plot, which is usually called an effect graph, it should be amply evident to readers that the patient-to-patient variability is the single largest source of systematic variation amongst the INR values observed in the study.

Table 15.9. A revised version of table 15.8 that includes the significant Machine × Reagent two-factor interaction

Term	SS	DF	MS	F	Significance level
Between Patients	16.230	58	0.280		
Within Patients					
Machine	1.581	2	0.790	239.6	<0.001
Reagent	2.598	1	2.598	787.5	<0.001
Machine × Reagent	0.087	2	0.044	14.3	<0.001
Residual	0.877	290	0.003		
Total	21.373				

Nonetheless, by designing their study carefully, and by involving a sufficiently large number of patients, the investigators have been able to uncover certain aspects of the complex process underlying INR measurement that are small, numerically, but important, scientifically.

Look carefully at the estimated mean INR values for each of the patients shown in figure 15.3. Regardless of which patient we consider, if the explanatory variables in the fitted regression model had involved only the main effect for Reagent rather than the main effect and the two-factor interaction of which Reagent was a part, i.e., X_1, X_4 and X_5, the two lines displayed on the effect graph would have been separated by the same distance, whether the machine was labelled 1, 2 or 3. But the two lines for patient 37 clearly aren't parallel. Neither is the pair of lines displayed for patient 27 since the separation between such lines is identical regardless of the patient considered. The change in mean INR that results when reagent 2 is used rather than reagent 1, i.e., the vertical distance between the solid and dashed lines associated with a particular patient, is evidently greater when INR values are measured on machine 3 than when either machine 1 or machine 2 are used. The same pattern is also evident in figure 15.1, where we can see that the amount by which mean INR changes when reagent 2 is used rather than reagent 1 also depends on which machine – 1, 2 or 3 – is used to measure INR. The effect of the change in reagent appears to be least when machine 1 is in use, and greatest on machine 3. This is a consequence of the significant two-factor interaction, and represents a physical or graphical interpretation of what this interaction means.

Even though ANOVA can be regarded as a specialized case of multiple linear regression, it is unlikely that one would choose to summarize the study results by providing a table of estimated regression coefficients and corresponding estimated standard errors. Instead, an ANOVA summary like table 15.9, and visual displays of the estimated effects, like figures 15.1 or 15.3, are usually presented. Although the final conclusions from this study are perhaps unsettling, scientifically, because of the subtle measurement effects that the data reveal, they highlight the fact that careful study design, combined with equally careful analysis of the resulting data, can provide important insights into complex processes.

Of course, a single example like the INR study that we have described hardly provides sufficient scope to address all the possible uses for analysis of variance methods. Nevertheless, we hope that this longer introduction has enabled readers to develop an appreciation of ANOVA that can serve them well in future encounters with this widely-used, well-developed set of statistical tools.

16

Data Analysis

16.1. Introduction

In the preceding chapters, we have discussed a number of statistical techniques which are used in the analysis of medical data. It has generally been assumed that a well-defined set of data is available, to which a specific procedure is to be applied. In this chapter, we adopt a broader perspective in order to address some general aspects of data analysis.

There is a necessary formalism to most statistical calculations which is often not consistent with their application. While the formal properties of statistical tests do indicate their general characteristics, their specific application to a particular problem can require adaptation and compromise. Data analysis, perhaps, is as much an art as it is a science.

Experience is the only good introduction to data analysis. Our aim, in this chapter, is to highlight a few principles with which the reader should be familiar. These should promote a more informed reading of the medical literature, and lead to a deeper understanding of the potential role of statistics in personal research activity. Where possible, we will use examples for illustration although, since they are chosen for this purpose, these examples may be simpler than many genuine research problems. Also, any analysis which we present should not be considered definitive, since alternative approaches may very well be possible.

16.2. Quality Data

'Garbage in, garbage out' is an apt description of the application of statistics to poor data. Thus, although it may be obvious, it is worth stressing the importance of high-quality data.

If information is collected on a number of individuals, then it is critical that the nature of the information be identical for all individuals. Any classification of patients must follow well-defined rules which are uniformly applied. For example, if a number of pathologists are classifying tumors, there should be a mechanism to check that the classification is consistent from one pathologist to another. This might involve a re-review of all slides by a single pathologist, or selected cases could be used as consistency checks. Formal statistical methods to examine data collected for the evaluation of such consistency are discussed in chapter 23. Here, however, we simply emphasize that identifying effective methods for primary data collection should be an important objective.

Of course, it is possible that data may be missing for some individuals. Provided a consistent effort has been applied to collect data, allowance for the missing data can frequently be made in a statistical analysis. Even then, however, if there are observable differences in a response variable between individuals with and individuals without particular information, any conclusions based only on those individuals with available data may be suspect. Considerable efforts have been directed towards developing methods to deal appropriately with missing data, but describing such methods is beyond the scope of this book. These techniques typically depend on assumptions that often cannot be verified. Therefore, the most effective way of dealing with missing data is to devote considerable effort to ensure that the amount of missing data is minimized.

Two major types of data collection can be identified; we shall call the two approaches retrospective and prospective. Retrospective data collection refers to data that were recorded at some previous time and subsequently are selected to be used for research purposes. The quality of retrospective data is often beyond the control of the investigator. Good detective work may provide the best information available, but what is available may vary widely from individual to individual. For retrospective data, classification frequently must be based on the greatest amount of information which is available on all patients. For example, in the 1970s it was shown that prior blood transfusions were associated with a poorer prognosis for aplastic anemia patients undergoing bone marrow transplantation. Patient records from the time prior to their arrival at the transplant center contained varying details on transfusion histories. As a result, early studies were necessarily limited to a simple binary classification indicating whether or not any blood transfusions had been used, even though, for some patients, the number of units of blood transfused could be identified.

Prospective data collection generally occurs in a well-designed study. In such a situation, specified information is identified to be of interest, and this

Table 16.1. Some information collected by questionnaires from 180 pregnant women

1	Patient number
2	Back pain severity
	(0) 'nil'
	(1) 'nothing worth troubling about'
	(2) 'troublesome, but not severe'
	(3) 'severe'
3	Age of patient (years)
5	Height of patient (meters)
6	Weight of patient at start of pregnancy (kg)
7	Weight of patient at end of pregnancy (kg)
8	Weight of baby (kg)
9	Number of children by previous pregnancies
10	Does the patient have a history of backache with previous pregnancy?
	(1) 'not applicable'
	(2) 'no'
	(3) 'yes, mild'
	(4) 'yes, severe'
13	Does walking aggravate back pain? (no/yes)

information is collected as it becomes available. The problem which arises in this type of study is one of ensuring that the information is, in fact, recorded at all, and is recorded accurately. In large collaborative studies this is a major concern, and can require considerable staff over and above the necessary medical care personnel.

Some additional aspects of data collection will be mentioned in chapter 18, which discusses the design of medical studies. In the rest of this chapter, only analyses of available data will be considered.

16.3. Initial or Exploratory Analysis

Before any formal statistical analysis can begin, the nature of the available data and the questions of interest need to be considered. In designed studies with careful data collection, this phase of analysis is simplified. It is always wise, however, to confirm that data are what they should be, especially if subsequent analyses involve computer manipulation of the data.

Table 16.1 presents a subset of some information obtained by questionnaires from 180 pregnant women [32]. The data were collected to study back pain in pregnancy and, more particularly, to relate the severity of back pain to

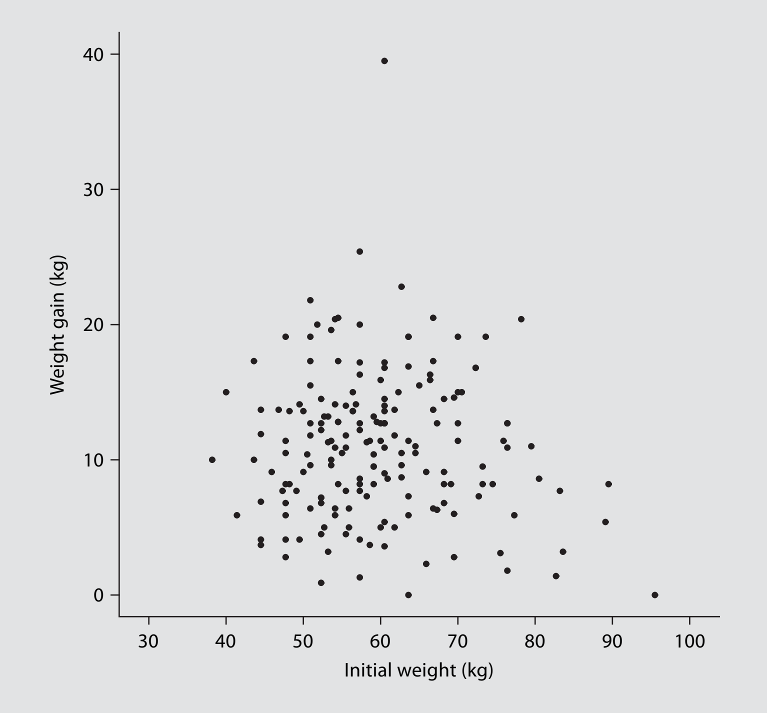


Fig. 16.1. A scatterplot of weight gain during pregnancy versus weight at the start of pregnancy for 180 women.

other items of information. The results of the questionnaires were kindly made available by Dr. Mantle to a workshop on data analysis sponsored by the Royal Statistical Society.

In this example, as in many medical studies, there is a clearly defined response variable which is to be related to other explanatory variables. If the response variable is not obvious, then it is important to consider whether such a distinction among the variables can be made, because it does influence the focus of the analysis.

The initial phase of an analysis consists of simple tabulations or graphical presentations of the available data. For example, figure 16.1 is a scatterplot of weight gain in pregnancy versus weight at the start of pregnancy. The most obvious feature of this plot is that one woman has a recorded weight gain of almost 40 kg, about twice that of the woman with the next largest weight gain. This is somewhat suspicious and should be checked. Such extreme values can seriously influence estimation procedures. Also, two women are identified as

Table 16.2. The responses to question 10 cross-tabulated by the responses to question 9 (see table 16.1)

Number of children by previous pregnancies	History of backache with previous pregnancies			
	0	1 ≡ not applicable	2 ≡ no	3/4 ≡ mild/severe
0	20	79	1	1
≥1	6	4	32	37

having a zero weight gain; the weights at the start and end of pregnancy recorded on their questionnaires were identical. Although this is not impossible, such data should also be checked. Since we cannot verify the available data, these three individuals will be omitted from subsequent analyses.

Table 16.2 is a table of the responses to a question about a history of backache in previous pregnancies and the number of children by previous pregnancies. This highlights another problem with the quality of the available data. Although the response to the question concerning a history of backache in previous pregnancies was supposed to be coded 1, 2, 3 or 4, 26 women coded the value 0. Also, two women with no children by prior pregnancies have recorded codes concerning back pain, and four women with previous pregnancies have recorded responses labelled not applicable. If possible, these responses should also be checked, but here we are forced to make the 'reasonable' assumption that the not applicable and zero codes for women with previous pregnancies correspond to no previous backache (coded 2), and we recode the responses for all women with no previous pregnancies as 1 (not applicable). To shorten the discussion, we shall ignore the possibility of miscarriage, etc.

One aim of these preliminary tabulations, then, is to clean up the data set. This can be a time-consuming operation in a large data set, where the consistency of many variables needs to be checked. However, inconsistencies must be resolved, and this sort of activity is an important component of data analysis.

Table 16.3 is an expanded version of table 16.2 after the miscoded responses have been revised. This table indicates that the majority of the women have had no children, or at most one child, by a previous pregnancy. For the few women with more than one child by previous pregnancies, the degree of back pain does not depend strongly on the number of pregnancies. Therefore, without performing any formal statistical tests, we might conclude that there is little to be gained from a detailed study of the number of children by previous pregnancies. Initially, formal analysis procedures may therefore be restricted to considering the simple binary classification of parous and nulliparous women.

Table 16.3. The revised responses to question 10 cross-tabulated by the responses to question 9 (see tables 16.1, 16.2)

Number of children by previous pregnancies	History of backache with previous pregnancies			
	not applicable	no	mild	severe
0	101	0	0	0
1	0	28	19	5
2	0	6	3	3
3	0	3	4	0
4	0	3	0	0
5	0	1	2	0
6	0	0	0	1
7	0	1	0	0

We will not discuss any additional exploratory investigation of these data. Nevertheless, exploratory analysis is important, and we hope some appreciation for this aspect of data analysis has been conveyed.

16.4. Primary Analysis

In many medical studies, there are clearly defined questions of primary interest. In a clinical trial, for example, the treatment comparison is the main purpose of the trial; any additional information is of secondary importance, or has been collected to aid in making a valid treatment comparison. We will assume that, in the back pain example we have been discussing, the primary purpose was an initial study of the influence on backache of unalterable factors such as age and previous pregnancies. This suggests that adjustment may be required for other factors such as weight gain during pregnancy. Regression models are frequently a useful method of analysis in such a situation.

In this study, the response variable, back pain, is of a type we have not previously discussed in the context of regression models. It is discrete, with four categories, and is naturally ordered. Regression models for such data do exist, extending, in principle, the ideas of logistic regression. However, the primary purpose of the analysis may not require the use of such a specialized technique. More important yet, we should consider whether the data warrant a highly sophisticated treatment. Reaction to pain is likely to be very variable among individuals. Because of this, it may not be sensible to use a method of analysis

Table 16.4. Current back pain severity versus a history of backache with previous pregnancies

History of backache with previous pregnancies	Current back pain severity				Total
	none	little	troublesome	severe	
Not applicable	8	56	28	9	101
None	5	14	14	9	42
Mild	0	7	13	8	28
Severe	0	3	5	1	9
Total	13	80	60	27	180

which places importance on the distinction between the back pain categories ‘nil’ and ‘nothing worth troubling about’. Similarly, the distinction between the upper two levels, ‘troublesome’ and ‘severe’, may not represent reliable data. For the primary purpose of the study, therefore, let us divide the back pain variable into two categories which represent the upper and lower two levels of response, assuming that this distinction will be realistic and meet the needs of the analysis. Logistic regression then becomes a natural choice for the method of analysis.

There are a variety of approaches to the use of logistic regression and the identification of important covariates which should be included in any model. Table 16.4 suggests that there is a relationship between a history of back pain in pregnancy and pain in the pregnancy under study. This variable would likely be included in a model. The inclusion of such variables in a regression model was previously discussed in chapter 15. In this case, to include the information represented by this categorical variable with four levels in the logistic regression model will require three binary covariates. Here, we will let the baseline category correspond to nulliparous women. The three binary covariates are then used to compare women in the three pain categories of none, mild and severe to the nulliparous group. The age of the patient is also a covariate of interest, and would be considered for inclusion in the model.

Table 16.5 presents the results of a logistic regression analysis which incorporates these two variables. The model indicates that the historical covariates are associated with current pain. Notice that the variable comparing parous women with no history of backache to nulliparous women is the least significant of the three historical comparison covariates and has a considerably smaller coefficient than the other two historical variables. If the coefficients for these three variables were comparable, we might suggest using a single vari-

Table 16.5. The results of a logistic regression analysis relating back pain in pregnancy to a history of backache in previous pregnancies and age

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
a	–0.61	0.23	–	–
Age <20 years	0.13	0.47	0.28	0.78
Age >30 years	0.14	0.40	0.35	0.73
History of backache in previous pregnancies				
None	0.64	0.39	1.64	0.10
Mild	1.64	0.50	3.28	0.001
Severe	1.23	0.75	1.64	0.10

able comparing parous and nulliparous women. Since the coefficients for the variables coding mild and severe pain in previous pregnancies are comparable, it suggests that the important predictive distinctions would be between women who are nulliparous, parous with no history and parous with some history of backache in previous pregnancies.

The coefficients for age are not significant in table 16.5, indicating that age is not related to back pain in pregnancy after adjustment for the historical backache information. In fact, age is not important, even if examined alone.

Conventional wisdom would suggest that weight gain in pregnancy will influence back pain. As is frequently the case, however, there are a variety of ways that weight gain could be introduced into the logistic regression model. Consider the three possibilities of actual weight gain, weight gain as a fraction of initial weight and actual weight gain divided by height. The last two variables attempt to take into account the influence of physical characteristics of a woman in carrying a child.

Table 16.6 records the logistic regression coefficients for each of these variables when added to the model specified in table 16.5. Also recorded are the coefficients when all three variables are added simultaneously. The most significant variable appears to be actual weight gain divided by height.

On the basis of these calculations, it would be customary to include actual weight gain/height and the historical pain variables in the logistic regression analysis presented in table 16.7. In a paper, table 16.7 might appear as a summary of the logistic regression analysis. The means of arriving at this model has been entirely suppressed, but this should influence our interpretation of the stated p-values.

Table 16.6. The logistic regression coefficients and estimated standard errors for weight gain variables added to the logistic model specified in table 16.5

Covariate	Added singly		Added jointly	
	estimated regression coefficient	estimated standard error	estimated regression coefficient	estimated standard error
Actual weight gain	0.07	0.03	−0.58	0.34
Fractional weight gain	3.91	1.83	−5.18	6.17
Actual weight gain/height	0.13	0.05	1.20	0.62

Table 16.7. A logistic regression analysis of the 'final' model for back pain in pregnancy

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
a	−0.49	0.44	–	–
History of backache in previous pregnancies				
None	0.73	0.39	1.87	0.06
Mild	1.76	0.49	3.59	<0.001
Severe	1.41	0.75	1.88	0.06
Actual weight gain/height	0.13	0.05	2.60	0.01

The variables relating to historical pain are straightforward, and the associated p-values are indicative of their relationship to back pain. However, the variable actual weight gain/height was chosen from a pool of other possible covariates precisely because it had a low p-value. This is one version of something called the 'multiple comparisons problem'. This issue frequently arises when regression techniques are used, and the reader should be aware of it. Non-technically, the multiple comparisons problem can be described as statistical testing of effects suggested by the data. In any data set there will be peculiarities, and if one searches long enough, they can be found.

As a simple illustration of this problem, consider the averages of some variable in four groups of individuals. If the highest average is compared with the lowest average using a simple t-test, then the p-value will be artificially low because there must be a highest value and a lowest value, and even if there are no real differences between the groups, the expected difference between the observed highest and lowest values must be positive. Perhaps this illustration is a statistical counterpart of the aphorism 'There is nothing you can't prove if your outlook is only sufficiently limited'.

In our analysis of the back pain data, the inclusion of a weight gain variable is important principally because it establishes an independent effect for the historical information on back pain in previous pregnancies. From this perspective, the choice between roughly comparable variables is not critical. No great importance should be attached to the form of the variable unless it is clearly more useful for predictive purposes than other choices in a variety of studies.

These general reservations concerning the uncritical interpretation of p-values in the context of multiple comparisons should be kept in mind, although it is difficult to formalize them. In §16.6, some formal consideration will be given to the same problem in a different context.

Although we have not presented all the details here, the logistic regression analysis will allow us to answer the primary questions of the back pain study. The experience of previous pregnancies is important in predicting back pain in pregnancy, while age is not important. This conclusion remains valid, even after adjustment for other variables which may also influence back pain.

16.5. Secondary Analyses

To answer the primary questions of a study, a degree of conservatism is usually wise. Often, the analyses undertaken should have been specified before the data were collected. The assumptions of any statistical model used should be consistent with the observed data, or seen as convenient simplifications which help to summarize the data without affecting the major conclusions.

After this phase of the analysis has been completed, additional data analysis is often undertaken. The distinction between primary and secondary analyses is not well defined, except perhaps in the case of clinical trials, and frequently may not be made at all. The reason we emphasize it here is to convey the notion that, in some settings, it is wise to downplay the importance of formal statistical tests.

For example, consider the back pain study which we discussed in §§16.3, 16.4. Factors which influence whether a woman with pain will call it severe or not may be of interest. Although this judgement may be quite subjective and variable, it is appropriate to look at data of this type.

For the 84 women in the back pain study who reported troublesome or severe pain, table 16.8 presents a logistic regression analysis with covariates which distinguish between these two categories. The covariates which dis-

Table 16.8. The results of a logistic regression analysis of severe versus troublesome pain

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
a	−1.51	0.39	–	–
Age <20 years	−0.22	0.82	0.27	0.79
Age >30 years	1.39	0.55	2.53	0.01
Walking aggravates pain	1.40	0.62	2.26	0.02

criminate between no pain and some pain, namely a history of backache and weight gain, do not discriminate between the upper two pain categories. Rather, the important covariates are age and an indication of whether walking aggravates pain. The effect of walking is only relevant to this logistic model since the patient must experience pain in order to aggravate it. Medically, one might speculate that this variable is a proxy for walking with bad posture, and plan to address this question in a later study. The fact that age plays an apparent role in the logistic model described in table 16.8, but not in the primary model, is somewhat surprising. Unless there is some reason to expect such a discrepancy, it would be wise not to stress the importance of an age effect until it could be confirmed in a subsequent study.

The general attitude of reservation which is expressed in the preceding paragraph is appropriate because if one continues to look at subsets of the data, then it is likely that something interesting will be found. As a hypothesis-generating activity, that is, to suggest future research questions, secondary analysis is valuable. However, the nature of this activity does undermine the probabilistic ideas of formal significance tests; therefore, the interpretation of secondary analyses should be treated with some caution.

When secondary analyses identify quite marked effects or relationships, then there is no reason they should not appear in published research. Other researchers are then able to consider the possibility of observing similar findings in related situations. However, the nature of the analysis which led to the findings should be made clear. Finally, in reading published papers, it is wise to consider whether the findings which are reported suggest that some discounting of reported significance levels would be prudent.

Table 16.9. Data concerning the presence of the HLA allele B8 in a diabetic and a control population

	B8 present	B8 absent
Diabetics	30	47
Controls	28	113

16.6. A Formal Discussion of Multiple Comparisons

In §16.4, we introduced the problem of multiple comparisons. Our discussion of this issue was necessarily general because the problem is often difficult to formalize. In some situations, however, specific allowance can be made for multiple comparisons.

Consider a sample of 77 diabetics and 141 normal individuals for whom HLA typing is available. More specifically, we will assume that, for each individual, the presence of alleles B7, B8, B12, and B15 at the B-locus of the human leukocyte antigen system on chromosome 6 can be detected.

Table 16.9 records presence or absence data for allele B8 in the two populations. A χ^2 test of independence leads to an observed value for the test statistic of 8.35 and a p-value of 0.004.

What has been suppressed in this presentation of the data is the fact that B8 is one of four alleles that were examined, and a similar 2 × 2 table for each allele could be produced. Table 16.9 presents the 2 × 2 table that leads to the most significant p-value. Once again, we have a situation where a significance test has been chosen on the basis of observed data.

If the number of comparisons which were undertaken can be counted, then a theorem in theoretical statistics due to Bonferroni suggests that the p-value for each comparison should be multiplied by the total number of comparisons undertaken. Hence, an adjusted p-value for table 16.9 would be $4(0.004) = 0.016$.

The appropriate application of the Bonferroni adjustment is not always clear, although it is a useful rule to keep in mind. When the results of such an adjustment are to be used to justify important conclusions, it would be prudent to seek statistical advice.

Another useful approach when multiple comparisons are a problem is known as the global test. Consider, for example, the relationship between previous backache and current back pain in pregnant women. Ignoring the effect of weight gain, table 16.10a cross-tabulates backache in previous pregnancies

Table 16.10. Cross-tabulations of current back pain in pregnancy versus backache in previous pregnancies: (**a**) full table; (**b**) collapsed table

a Full table

Current back pain	Backache in previous pregnancies				Total
	no prior pregnancy	none	mild	severe	
None or mild	64 (52.18)[1]	19 (21.70)	7 (14.47)	3 (4.65)	93
Moderate or severe	37 (48.82)	23 (20.30)	21 (13.53)	6 (4.35)	87
Total	101	42	28	9	180

b Collapsed table

Current back pain	Backache in previous pregnancies			Total
	no prior pregnancy	none	mild or severe	
None or mild	64 (52.18)[1]	19 (21.70)	10 (19.12)	93
Moderate or severe	37 (48.82)	23 (20.30)	27 (17.88)	87
Total	101	42	37	180

[1] The values in parentheses are expected numbers if the row and column classifications are independent.

by the two categories of current pain used in the logistic analyses which we discussed in §16.4.

One test of the hypothesis that the historical backache classification is independent of the current pain category is the χ^2 test for rectangular contingency tables discussed in §4.3. For this test, the observed value of the test statistic is 15.44 on $1 \times 3 = 3$ degrees of freedom. The 0.01 and 0.001 critical values of a χ^2_3 distribution are 11.345 and 16.268, respectively, indicating that the p-value for this test of the hypothesis of independence lies between 0.01 and 0.001. There is, therefore, strong evidence of an association between backache in previous pregnancies and current back pain. In reaching this conclusion, we have treated 'no prior pregnancy' as a separate category with respect to the classification of backache in previous pregnancies. If recurring backache in pregnancy was the primary question of interest, then a table that was restricted to women in the data set with at least two pregnancies would be appropriate.

Table 16.11. The results of a logistic regression analysis of diabetic status classified by HLA-B alleles

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
a	−0.72	0.27	–	–
B7	−0.65	0.40	1.63	0.10
B8	0.77	0.34	2.26	0.02
B12	−0.51	0.36	1.42	0.16
B15	0.71	0.37	1.92	0.05

Although, in this example, there is strong evidence for a relationship between the two tabulated variables, it is often possible to increase the significance of a rectangular contingency table by choosing to pool certain categories. If the grouping is performed because it appears from the data that a smaller p-value can be obtained, then the observed p-value needs to be discounted. For example, if we group the mild and severe pain categories, as is done in table 16.10b, then we can generate an observed value for the test statistic of 14.90 on two degrees of freedom, with a corresponding p-value which is less than 0.001. In general, investigators should be wary of grouping categories in order to strengthen findings, especially if the logic of the grouping is not natural, scientifically.

Table 16.11 presents the results of a logistic regression analysis of diabetic status for the 218 individuals discussed at the beginning of this section. The model involves explanatory covariates which are binary, indicating the presence or absence of HLA-B alleles B7, B8, B12 and B15. For example, the covariate labelled B7 is equal to one if allele B7 is present on at least one of an individual's two chromosomes, and is equal to zero otherwise.

The definition of these binary explanatory covariates reflects the available genetic information. The resulting analysis involves an implicit assumption that not all individuals will have one of the specified alleles. This assumption will be true if not all alleles at a particular locus can be detected, or if attention is restricted to a small set of alleles. For the purposes of illustration, we have adopted the latter perspective in this particular case. Finally, the logistic regression analysis summarized in table 16.11 is based on HLA typing procedures which do not allow the determination of homozygosity; consequently, this information is not coded in the explanatory covariates. More detailed modelling of genetic effects is possible; however, we do not intend to pursue it here.

From table 16.11, we see that allele B8 appears to have the strongest association with diabetic status. If we omit the other covariates, then in the resulting single covariate model involving variate B8 we would observe the same p-value of 0.004 as we saw previously in table 16.9. It would be inappropriate to report this p-value, since the test was suggested by the less specific analysis of table 16.11.

Even the results of table 16.11 need to be interpreted carefully. The four covariates in table 16.11 jointly classify the study subjects according to all coded allelic information. However, each covariate also represents a simple classification of the individuals on the basis of a single allele. In the logistic regression analysis, one of the covariates must have the lowest p-value. Unless this particular classification was of prior interest, the observed p-value should be interpreted with caution.

With sets of classification variables such as the one we have described, a conservative approach to regression modelling involves a global test of the significance of the classification scheme. Formally, this is a test of the hypothesis that all the regression coefficients associated with the classification variables are zero. We have not previously discussed tests of this kind in the context of logistic regression models, but we can do so, briefly, at this point.

Corresponding to any estimated regression model, there is a number called the log-likelihood. If we have a regression model with a log-likelihood of L_1 and add k new covariates to the model, a new log-likelihood L_2 will result. If the null hypothesis that *all* the new covariates are unrelated to the dependent or response variable is true, the test statistic $T = 2(L_2 - L_1)$ is distributed, approximately, as a $\chi^2{}_k$ random variable. A computer program is usually necessary to perform this global test, and a statistician should be consulted to carry out the detailed calculations and assist in interpreting the results of the test. This procedure is sometimes called a deviance test since it is becoming common to refer to twice the log-likelihood of a model as the deviance. Global tests were discussed in chapter 15 in the context of analysis of variance.

For the HLA data, the model presented in table 16.11 would be compared with the simple one-parameter model which omits the four HLA-B classification variables. This leads to an observed value for T of 19.8. The 0.001 critical value for a χ^2_4 variate is 18.465; therefore, the p-value for the global test is less than 0.001. Since this global test of the HLA-B classification variables is significant, we can be more confident that the significant p-values for individual variables are not spurious.

16.7. False Discovery Rates

Global tests and Bonferroni-corrected significance levels have a long history of use in relation to the problem of multiple comparisons. Prompted by new examples of very large and complex data sets, particularly but not solely in the area of genetics, other methods of addressing this problem are being developed. An approach suggested by Benjamini and Hochberg [33] that is based on the concept of *false discovery rates* has already demonstrated its usefulness.

If there are K hypotheses to test, then an unknown number, m, of these will be true, and K – m will be false. Based on appropriate tests of significance and a particular set of data, T of the m true hypotheses will be rejected, resulting in false positive outcomes. Likewise, F of the K – m false hypotheses will be rejected, constituting false negative outcomes. Conceptually, T and F are random variables. The false discovery rate (FDR) is defined as the expected value of T/(T + F). Thus, the focus of this approach is the number of statistically significant 'findings' that are actually false. These erroneous findings are controlled by choosing a significance level threshold at which hypothesis tests are deemed to be significant such that the FDR is less than some pre-specified level, e.g. 10%.

We will not discuss the use of FDRs in any more detail but simply want to make the reader aware of this potentially useful concept.

It is impossible to present a comprehensive discussion of the complexities of data analysis, and the possible caveats which must be considered in presenting the results of a statistical analysis. However, we hope that this brief introduction will alert readers to some of the more important considerations.

17

The Question of Sample Size

17.1. Introduction

In chapter 18, we intend to introduce several of the important issues that arise in the design of clinical trials. The question of sample size is a technical consideration comprising one aspect of the general problem of design. Although, in general terms, it is difficult to specify how many subjects are required to make a clinical trial worthwhile, to embark on any study without considering the sample size which is adequate is unwise, and may even be unethical. It is not appreciated widely enough that failure to detect a treatment difference may often be related more to inadequate sample size than to the actual lack of a real difference between competing therapies. In the final analysis, studies with inadequate sample sizes serve only to confuse the issue of determining the most effective therapy.

Sample size calculations are frequently complicated; therefore, we do not intend to describe the actual details in depth. Instead, we propose to highlight those aspects of the subject which are common to all situations, describing these features in non-technical terms. To illustrate the basic concepts in a more practical setting, we discuss the determination of appropriate sample sizes for two different hypothetical studies in §17.3. And, in the final section of the chapter, we point out some of the hazards associated with small clinical trials.

17.2. General Aspects of Sample Size Calculations

It is important to realize, right from the start, that sample size calculations will always be approximate. It is clearly impossible to predict the exact outcome of any particular clinical trial or laboratory experiment. Nevertheless, the importance of sample size calculations is demonstrated by the fact that they provide information about two important design questions:

(1) How many subjects should participate in the intended experiment?

(2) Is this study worth doing if only n subjects (a fixed number) participate?

Clearly, both questions enable an investigator to evaluate the study or experiment critically, and to decide whether to proceed as planned, or perhaps to revise the overall design. In certain cases, the wisest decision may be not to initiate the study because the likelihood of demonstrating the desired effect, using the available resources of personnel, funds and participants, is very small.

In chapter 16 we suggested that, in most studies, there will be a primary question which the researchers want to investigate. The calculations concerning sample size depend on this primary question and the way in which it is to be answered. Therefore, in order to answer either of the two questions posed above, an investigator needs to decide what sort of data will be collected and how these data are to be analyzed. These choices need not be rigidly determined; indeed, they never can be since almost every study will involve some unexpected features. All the same, since sample size calculations depend on the proposed method of analysis, some tentative assumptions are necessary. Already, it should be obvious to the reader why sample size calculations are only approximate.

For the sake of illustration, let us suppose that a clinical trial is being planned. The chief purpose of the trial is to compare the effectiveness of an experimental treatment with the current standard procedure, called the control treatment. Without being too specific, we can state that an answer to the primary question can be expressed in terms of a clinically relevant treatment difference. For example, this difference might be the reduction in mortality experienced by patients receiving the experimental treatment. The study is being conducted to determine the degree to which the results are consistent with the null hypothesis of no treatment difference. Either a treatment difference will be demonstrated, or the data will be judged consistent with the null hypothesis.

An additional degree of artificiality is introduced into sample size calculations by the convention that failure to reject the null hypothesis is equivalent to concluding that the null hypothesis is true. This convention is inappropriate at the time of analysis, but convenient at the design stage.

The true situation in the study population concerning the null hypothesis, H_0, can never be known for certain. Also, with respect to the true situation, the researcher's final conclusions regarding the treatment difference may be correct or wrong. Table 17.1a concisely sums up the framework within which sample size calculations are undertaken.

Table 17.1. The hypothetical framework of sample size calculations: (**a**) correct and erroneous conclusions; (**b**) probabilities of correct and erroneous conclusions

True situation	Investigator's conclusion	
	H_0 True	H_0 False
a Correct and erroneous conclusions		
H_0 True	correct conclusion	false positive (Type I error)
H_0 False	false negative (Type II error)	correct conclusion
b Probabilities of correct and erroneous conclusions		
H_0 True	$1 - \alpha$	α
H_0 False	β	$1 - \beta$

In two of the four cases that could arise, the researcher will reach a correct conclusion. However, in the remaining two cases the researcher will be wrong, having reached either a false positive or a false negative conclusion. Traditionally, statisticians have called these erroneous outcomes Type I and Type II errors, and have represented the probabilities of these two outcomes by α and β, respectively (see table 17.1b). In our view, the terms false positive and false negative are more informative in the medical research setting.

If no treatment difference exists, then α is simply the probability of obtaining an unlikely outcome in that situation and therefore deciding that the data contradict the null hypothesis. This probability is precisely the threshold for the significance level of the data with respect to H_0, i.e., the p-value. The value of α which the investigator uses to conclude whether H_0 is true or false is usually regarded as fixed in advance.

If a real treatment difference exists, i.e., if H_0 is false, the probability that the investigator will correctly conclude that this is so is $1 - \beta$ (see table 17.1b). This probability depends largely on N, the total sample size, but also on the actual magnitude of the treatment difference. The reason for this dependence on total sample size is fairly simple. A larger sample contains more information about characteristics which are of interest, and hence facilitates more precise estimation of the true situation that obtains in the study population. Therefore, by increasing the sample size, we increase our ability to detect any real treatment difference which exists. This increased ability to determine whether H_0 is true or false is translated into an increased probability, $1 - \beta$, that a correct conclusion will be reached when H_0 is false. Since $(1 - \beta) + \beta = 1$, if $(1 - \beta)$ increases, then the false negative probability, β, necessarily decreases. Therefore, increasing the sample size can also be viewed as a means of decreasing the false negative rate when a real treatment difference exists.

We are now in a position to state, in precise terms, the two questions about any study which sample size calculations will answer:

(1) If the probability of a false positive conclusion is fixed at α, what total sample size, N, is required to ensure that the probability of detecting a clinically relevant difference of given magnitude δ is 1 – β?

(2) If the probability of a false positive conclusion is fixed at α, and a specific sample size, N, is employed, what is the probability, 1 – β, that the study will detect a clinically relevant difference of given magnitude δ?

Notice that to answer the first question we must specify sample sizes, N, which correspond to prescribed triples of α, δ and 1 – β. On the other hand, to answer the second question we must determine values of 1 – β which correspond to specific values of α, δ and N. In either case, the answer would not be a single value, but rather a table, or perhaps a graph, of sample sizes, N, or probabilities, 1 – β.

Of course, other considerations besides the method of evaluation and the clinically relevant treatment difference will affect the determination of adequate sample size. These include the relative sizes of the treatment groups, possible dropout rates in these groups and the thorny problem of treatment noncompliance. Unfortunately, a researcher may have no effective means of controlling some of these factors which can seriously affect the anticipated outcome of the study. In addition, any failure to observe uniform standards in evaluating patient characteristics or treatment outcomes may increase overall variability in the study. The inevitable consequence will be a reduced ability to detect any real treatment difference which exists.

It is beyond the scope of this book to give many of the details involved in performing sample size calculations. The time spent consulting a statistician, or undertaking additional reading, on this aspect of study design will be seen, we hope, as time well spent. To prepare the reader for such activities, in the following section we discuss two hypothetical studies and describe the aspects of each which would be considered in evaluating the required sample size.

Comments:

(a) Since α and β both represent probabilities of making an erroneous decision, in the best of all possible worlds we would like both α and β to be close to zero. Unfortunately, if α is decreased without changing the total sample size, N, then β necessarily increases. Conversely, if β must decrease without changing N, then α necessarily increases. Only by increasing the sample size can a simultaneous reduction in both α and β be achieved.

(b) Typically, the value of α is fixed by the experimenter, since α is the p-value at which the study outcome will be regarded as statistically significant. In this case, β will decrease as the total sample size increases. The decision regarding an adequate sample size for a given study will necessarily be a compromise, balancing what can be achieved, statistically, with a sample size that is practical.

(c) The probability, $1 - \beta$, of detecting a specified difference, δ, is called the power of the study. A powerful study is one with a high probability of detecting an important treatment difference. It has been proposed that if a study fails to reject the null hypothesis, then it is important to state the power of the study. Although this proposal would aid in the interpretation of a study's conclusions, we believe it is an inappropriate use of power calculations. Sample size (or power) calculations are relevant to study design, not analysis. At the analysis stage, the results of significance tests and statistical estimation are relevant. If a confidence interval is provided for an estimated treatment difference, then power calculations will furnish no additional information. In addition, the assumptions inherent in power calculations are generally more restrictive than those required at the analysis stage.

The two examples of sample size calculations discussed in the following section are very simple. They only serve to illustrate how some of the essential features that are involved in the preparation of sample size estimates are evident in these particular applications. More generally, detailed sample size calculations can often be provided by modern statistical software packages. However, readers who seek the more intangible benefits that derive from a careful evaluation of the feasibility of a proposed study would be wise to consult a statistician rather than to rely solely on the uncritical use of software packages.

17.3. Two Examples of Sample Size Calculations

17.3.1. A Comparison of Proportions

Several researchers are proposing to conduct a clinical trial to evaluate the effectiveness of the β-blocker, metoprolol, in reducing mortality following acute myocardial infarction. It has been decided that the outcome of interest in the study is to be death within 90 days of the initial attack. While many of the details of the study protocol have yet to be determined, the study collaborators have decided to conduct a double-blind, randomized, controlled trial using a suitable placebo. With respect to the primary purpose of the study, the data will be analyzed using the methods for 2×2 contingency tables.

Effectively, the study is designed to compare the proportions of deaths within 90 days (90-day mortality rates) observed in the two study groups. The researchers anticipate that the mortality rate for metoprolol will be considerably lower than the corresponding value for the placebo. How large should the total sample size for the study be?

In order to carry out sample size calculations, the consulting statistician needs some additional details regarding the study. The researchers will have to estimate the mortality rate which they expect to observe in the placebo group; in addition, the clinically relevant difference between the two mortality rates which they want to be able to detect is also needed. If unequal group sizes are to be used, the fraction that each group represents of the total sample size must be specified. Finally, the investigators must determine the value of α, the probability of a false positive result, which they are willing to accept. Once these items have been specified, the statistician will be able to generate a table of sample sizes and the corresponding values of β, the probability of a false negative result.

After some discussion, the researchers report that the expected mortality rate in the placebo group is about 0.20 and the clinically relevant difference is a reduction in the mortality rate to 0.10 or less. This is the treatment difference that they want to be able to detect. The study participants will be randomized in equal numbers to the two treatments and the traditional α-value of 0.05, i.e., $p = 0.05$, will be used to evaluate the statistical significance of the study.

In many cases, a statistician is able to refer to statistical tables for sample size calculations which help to answer a client's query. For the particular situation described above, table 17.2a, which indicates the required net total sample size, N, and the corresponding probability of a false negative, β, might be prepared.

Notice that the calculations specify net total sample size. If provision is to be made for patient dropout and treatment noncompliance, the necessary sample sizes will be larger than those indicated in the table.

17.3.2. An Analysis of Survival Data

In this hypothetical study, the researcher is proposing to conduct an experiment concerned with death due to vaginal cancer in rats insulted with the carcinogen DMBA. Two groups of equal size will have differing pretreatment regimens. The rats will be observed for a specified period. The primary purpose of the study is to determine whether survival depends on the pretreatment regimen.

Observations on rats which die of causes unrelated to the application of the carcinogen and are free of tumor at death, or rats which simply have not developed tumor at the time of data analysis will be censored observations on the time of death from vaginal cancer. To analyze the experimental data, a log-rank test (see chapter 7) for the treatment difference will be used.

Table 17.2. Two sample size tables for hypothetical studies: (**a**) comparing proportions; (**b**) analyzing survival data

a Comparing proportions

N	100	200	300	400	500	600	700
β	0.71	0.48	0.31	0.19	0.11	0.06	0.04

b Analyzing survival data

Survival rate in the poorer group at the time of analysis	Survival advantage enjoyed by the better group	
	0.05	0.10
0.05	497 (661)[1]	174 (232)
0.10	963 (1289)	295 (395)
0.15	1415 (1894)	406 (544)

[1] Sample size required to yield $\beta = 0.20$ (0.10).

In order to prepare some sample size tables for this particular study the following are required: the proportion of rats in each treatment group which are likely to be free of tumor at the time the experimenter decides to terminate the experiment and α, the probability of a false positive result which will be used to determine the statistical significance of the log-rank test. With some difficulty, the investigator eventually estimates that the survival rates in the two groups at the time of analysis are likely to range from 0.05 to 0.15 for the poorer group, with the other treatment group possibly enjoying an advantage of 0.05 to 0.10. The usual probability of a false positive conclusion, namely $\alpha = 0.05$, will be used to evaluate the log-rank test.

Based on these values and sample size tables for the log-rank test produced by Freedman [34], the statistician draws up table 17.2b. Entries in the table specify the total sample size which is required to yield 0.20 probability of reaching a false negative conclusion with respect to the indicated survival advantage (treatment difference). The numbers in parentheses correspond to a false negative probability level of 0.10.

The values given in the table presuppose that all rats either die of vaginal cancer or are observed free of tumor at the time of analysis. If an appreciable fraction of the original experimental group are likely to yield censored observations prior to the termination of the experiment, then the numbers given in table 17.2b will need to be increased to adjust for this loss in information regarding death due to vaginal cancer.

In both of the examples we have just considered, the ‘answer’ regarding sample size has not been a single number, but rather a table of values. This method of presentation highlights several of the features of sample size calculations which we discussed in §17.2. For example, sample size calculations are, by nature, approximate, since the results depend on tentative assumptions regarding the expected outcome of the study. Moreover, the actual sample size which a researcher uses will inevitably be a compromise between the values of α and β which are acceptable, and the cost of the study in terms of resources and time per participant. More important still, sample size calculations can convince researchers that when patient numbers are too limited, certain studies should be concluded before they begin.

17.4. Some Hazards of Small Studies

In preceding sections of this chapter we have suggested, without justification, that studies which involve only a small number of participants may not be worth doing. In this final section, we intend to provide the reader with sound reasons to be wary of conclusions which are derived from small experiments or clinical trials.

As a starting point, let us consider the following hypothetical situation which is discussed in Pocock [35]. Suppose the response rate for the standard treatment of a disease is 0.30. A number of new drugs are being developed and require evaluation in carefully-conducted clinical trials. Of course, not all the drugs will be more effective than the standard treatment. In fact, we will assume that 80% are no better than the standard treatment. The remaining 20% of new drugs will achieve a response rate of 0.50, and therefore represent a major advance in the treatment of the disease. The chief purpose of each clinical trial which may be held is to determine whether the drug being tested belongs to this latter category.

A number of clinical trials are held, worldwide, and in each trial one new drug is evaluated relative to the standard treatment. All trials are of the same total size, and the value of α, the probability of a false positive conclusion if the null hypothesis of no treatment difference is true, is the conventional 0.05. Each trial will be summarized as a 2 × 2 contingency table and analyzed accordingly.

The situation we have just outlined is admittedly simplistic. Nevertheless, we believe it can illustrate two of the major problems associated with small studies of any kind. Table 17.3 summarizes the expected outcome of the situa-

Table 17.3. Details of hypothetical clinical trials of five sizes illustrating some hazards of small studies

Size of trial	True response rates (experimental – standard)	Number of trials worldwide	Expected number of statistically significant trials ($p \leq 0.05$)	Percentage expected of true positives detected	Ratio of expected false positives to expected true positives
400	0.30–0.30	40	2	–	0.20
	0.50–0.30	10	9.9	99	
200	0.30–0.30	80	4	–	0.22
	0.50–0.30	20	17.9	90	
100	0.30–0.30	160	8	–	0.30
	0.50–0.30	40	26.4	66	
50	0.30–0.30	320	16	–	0.47
	0.50–0.30	80	34.0	43	
25	0.30–0.30	640	32	–	0.74
	0.50–0.30	160	43.0	27	

tion we have described for clinical trials of five different sizes – 400, 200, 100, 50 and 25 participants. For each of these five situations, a total of 20,000 participants are involved, worldwide.

The most important columns in table 17.3 are the latter three; these indicate the expected number of trials which would be statistically significant ($p \leq 0.05$), the percentage expected of true positives that would be detected, and the ratio of expected false positive to expected true positive trials. Notice, first, that as the trial size decreases from 400 to 25, the expected true positives detected decreases from 99 to 27%. Thus, trials involving few participants are clearly less capable of identifying the effective drugs than larger trials. Moreover, as trial size decreases, the ratio of expected false positive conclusions to expected true positives increases from 0.20 to 0.74. This reflects the fact that, since the value of α is usually fixed, a large number of small trials increases the likelihood that a positive conclusion is a false one.

But there are other problems which could easily arise if a particular study were to involve only a limited number of subjects. We will argue, in chapter 18, that randomization is advisable whenever its use is ethically defensible. To a statistician, randomization in a clinical trial is rather like the premium which a home owner pays annually for fire insurance coverage. If an important factor happens to be overlooked, inadvertently or unconsciously, during the planning of a certain trial, randomization should ensure that its effect is roughly similar in all treatment groups. However, the effectiveness of randomization depends on total sample size; randomizing a large number of participants is more likely to achieve the intended result.

Our discussion of the hazards of small studies has not been exhaustive. Nevertheless, we hope that the need for a cautious interpretation of such studies has been demonstrated. Some years ago, a number of countries began to require that cigarette packages should bear a warning message for users. Perhaps small studies deserve to be similarly distinguished.

18

The Design of Clinical Trials

18.1. Introduction

The main emphasis of this book is the analysis of medical data. The quality of data available for analysis clearly depends on the design used for its collection. In a medical trial, investigators must balance considerations of ethics, simplicity and good statistical practice, and it is often difficult to give anything more than general advice about the characteristics of a well-designed study. However, there are a number of good resources on the design of clinical trials which can be consulted for more detailed discussion.

In this chapter, we shall briefly present a few of the issues which are frequently discussed and comment on trial organization. We also explore the role of randomized treatment assignment in clinical trials in somewhat greater detail. The use of randomized trials has been the subject of considerable debate, and we feel it deserves some discussion here. Sections on intention-to-treat analyses, factorial designs and repeated significance testing provide a short overview of these aspects of trial design. Finally, we conclude the chapter with a brief introduction to the important topic of the sequential analysis of a clinical trial.

18.2. General Considerations

Perhaps the primary requirement for a good clinical trial is that it should seek to answer an interesting question. The choice of treatments to be compared and the patients on whom they are to be compared largely determine the practical importance of discovering whether the treatments differ.

Strict entrance requirements which generate a very homogeneous patient population facilitate precise treatment comparisons with a small number of patients. However, the results of a larger study with a more heterogeneous population would probably be more convincing to a practising physician.

A trial with two highly divergent treatments is simple and is likely to produce a result more quickly than a trial with two similar treatments, or one involving more than two treatments. This observation is an important one since, for various reasons, it is often tempting to stop a trial before conclusive results have been obtained. On the other hand, sophisticated designs frequently allow more comprehensive inferences to be deduced. It is also important to ensure that the intended treatments are acceptable to the clinicians who must enroll their patients into the trial. Therefore, in selecting treatments, a balance must be struck among these various factors.

The design stage of a clinical trial should also specify data collection procedures. The information which will be collected concerning each patient at entry into the study should be identified. These baseline variables can be used in the analysis of the trial results to adjust for patient differences in the treatment arms. Therefore, information which is gathered at entry should be related to the chosen endpoints or to potential side effects of treatment; this latter aspect is sometimes overlooked. Since collecting data on a patient at the time of entry into a study is generally easier than attempting to recover relevant baseline information at a later time, it is advisable to record as much baseline data as is feasible.

Collecting data on only a few endpoints will make follow-up easier, and also reduces the chance of serious bias due to differential follow-up among patients. At the same time, as much information as possible should be recorded concerning each endpoint of the study. The time until a certain event is observed is more informative than a mere record of its occurrence. All patients who enter a trial should be followed, even if they abandon the treatment protocol, since exclusion of these patients can introduce bias. Similarly, the treatment groups which are compared in the primary analysis should be groups based on the treatments which were originally assigned (see also §18.5), because this comparison reflects how the treatments will perform in practice. Of course, it may be of scientific interest to restrict a comparison to those patients receiving and tolerating treatment regimens, for example, but the more general comparison, based on assigned treatments, is usually more valuable in the long run. Note that, in order to avoid bias, treatment assignment should only occur after informed consent procedures.

Another point of frequent discussion concerns the stratification of treatment assignment by prognostic factors. The statistical methods which have been developed, such as regression models, reduce the need for precisely comparable treatment groups. It seems reasonable, however, to consider stratifying a trial on a few factors of known prognostic significance and to attempt partial balance on other factors via randomization. The effectiveness of the randomization in achieving this balance should be examined. Peto et al. [36] argue in favor of no stratification, but it is more cautious, and perhaps more convincing, to balance on major prognostic factors rather than to rely solely on sophisticated statistical analyses to adjust for imbalances in the treatment arms. While it is generally agreed that excessive stratification is complicated, often unnecessary, and may even result in poor balance if only a few patients are entered in each stratum, an alternative to stratified randomization does exist. The advent of widely available computing resources allows the use of a technique called *minimization.* Minimization aims to provide an effective randomization scheme when there are more than two or three prognostic factors on which stratification might be appropriate, and therefore the risk that stratification on those factors will lead to poor balance is no longer negligible.

The goal of minimization is not to ensure balance within each of the potentially many strata that are defined by all possible combinations of the relevant prognostic factors. Instead, it only aims to ensure that, when each prognostic factor is examined individually, there is appropriate balance between treatment assignments. The balance that has been achieved prior to randomizing a newly enrolled subject is examined, and the probability of assigning that subject to each of the various treatments offered in the trial is then specified for the new subject in a way that is likely to reduce any imbalance that may be present. The algorithm to specify the appropriate randomization probabilities is relatively complex and thus requires access to computer resources. It is this complexity which provides protection against selection bias.

There is debate concerning certain issues raised by the use of minimization, but we are not able to address those issues here. While it represents a method of treatment assignment which is being used increasingly, some caution about its routine adoption may be wise. For many trials, a moderate level of stratification may be sufficient, and easier to implement.

In the early design stage, the inferences which are to be drawn from the study should be identified. For example, suppose that a clinical trial of two adjuvant therapy regimens following surgery for breast cancer is being planned. The response variables which are of interest are remission duration and survival. The study protocol should specifically mention that remission duration and survival will be used to compare the two treatment regimens. In addition, the statistical procedure that will be used to analyze the results of the trial should be specified.

When the study has ended and the data are analyzed, it may be determined that the treatment A arm of the study had fewer metastases to the ovary than the treatment B arm, and this difference is statistically significant at the 5% level. Perhaps no other site of metastasis suggested there was a difference between the treatments, and the comparison of the two treatments on the basis of remission duration also indicated no difference. These results should not lead us to conclude that treatment A is better than B.

If ten sites of relapse were examined, then, because of the multiple comparisons problem which we discussed in chapter 16, it is not unlikely that one of the ten sites will, by accident, suggest there is a difference between treatments. If there was no reason, prior to the study, to suspect a treatment effect at a particular site of relapse, then the discovery of such an effect should be viewed with caution, especially if the designated principal comparison does not identify a treatment effect.

A major reason for specifying, in advance, the statistical procedure which will be used in the analysis is that it is possible to find perhaps ten different statistical tests which compare remission duration in two treatment groups. It might happen that one of these tests is just significant at the 5% level, while the rest suggest there is no significant difference. Such a 'search for significance' is entirely inappropriate; therefore, a reasonable test procedure should be specified before the study begins and used when the data are analyzed.

It may be that there is valid medical information in the unexpected results from a single relapse site, or that the statistical test indicating a treatment difference is particularly sensitive for the type of data produced by the study. If there is reason to suspect that this is the case, then the results should certainly be reported. How the results arose should also be reported, and it should be made clear that they need to be confirmed in other studies before being generally accepted. On the other hand, one can be much more confident about a result identified by a test which was specified prior to a detailed examination of the data.

18.3. Trial Organization

The previous section dealt with issues that arise in the design of clinical trials in a very general way. To implement an actual trial requires that attention be given to a myriad of details. In this section we comment briefly on the major aspects of trial organization. Our aim is simply to highlight key features, and we assume that interested readers who wish to undertake a clinical trial will both read more widely and hold discussions with researchers who have experience in running trials.

Specific considerations may arise if a trial is undertaken with an aim of gaining regulatory approval for a new treatment, as is common for new drugs developed by pharmaceutical companies. We do not attempt to discuss details concerning such trials here but, of course, the basic structure of all trials should be similar.

The bedrock of any clinical trial is the trial protocol. In this document the details regarding justification for the trial question, the details of the trial treatments, eligibility of patients, assignment of treatments to patients, primary outcomes, primary analyses and the monitoring of trial progress are specified, along with the sample size calculations that justify the expected size of the trial. Recently, some medical journals have adopted the policy of not publishing trial results unless a trial protocol has been officially registered prior to the beginning of the trial. In any event, it is the protocol that drives the day-to-day activities associated with the trial.

The trial protocol is also central to the submissions that are made to ethics committees which approve the implementation of the trial in various clinical jurisdictions. Considerable resources can be required to attain the necessary ethical approvals, especially if a trial aims to recruit patients from more than one centre. Increasingly, there are national and international guidelines about the running and reporting of trials. Demonstrated adherence to these requirements also requires resources.

Most trials will require a specific source of funds. A research grant proposal to support a trial will usually provide information similar to that found in a trial protocol, but will generally devote more attention to the justification for the trial, including a summary of available information on the proposed treatments. The latter will sometimes include a meta-analysis of related studies; chapter 20 provides a brief introduction to that subject. A grant proposal will usually include less clinical detail than does the protocol; however, adequate financial details will be required to support the request for trial funding. The submission of a grant proposal may precede the completion of a draft trial protocol; nonetheless, the basic outline of the protocol must be in place in order to achieve funding for the trial.

The most common organizational structure for a trial is to have three primary committees. The first is often called the Trial Management Committee or Team. This group of individuals is usually headed by the principal investigator(s) for the trial, and is charged with the day-to-day running of the trial. The Trial Steering Committee, on the other hand, is chaired by an individual who is independent of the investigators who designed and are implementing the trial, in order to provide independent oversight of the study. The Steering Committee will have other independent representation, usually including statistical expertise, and often including lay individuals from either the general public or disease interest groups. Trial investigators may also sit on this committee. The Trial Steering Committee is vested with primary responsibility for the ethical running of the trial. The final committee is usually the Data Monitoring Committee (DMC). This is a group of individuals who are entirely independent of the trial, and who are charged with monitoring the trial from an ethical perspective. The responsibilities of the DMC include ensuring that the trial is recruiting patients in a timely manner, being aware of adverse events associated with trial treatments and, when necessary, having access to the accumulating evidence concerning possible efficacy differences among trial treatments. The minimal requirement for such a committee is to have specific expertise in the clinical area of the trial, statistical expertise to interpret trial data, and general experience in the running of clinical trials. Specific ethical or legal expertise and lay participants may be included in such a committee, but there is wide variation from one trial to another. The DMC is advisory and reports to the Trial Steering Committee.

We have not done justice to the complexities of trial organization in this very brief discussion, but reiterate that there are many specialized sources of additional information for the interested reader.

18.4. Randomized versus Historical Controls

A randomized clinical trial is generally regarded as the 'gold standard' for a clinical investigation. Nevetheless, there can be questions concerning the use of such trials. One of the major concerns is often the ethical problem of allowing a random event to determine a patient's treatment.

We do not intend to summarize the various issues which have been discusssed with respect to randomized clinical trials. Arguments for and against their use have been advanced in the past, and interested readers can consult references 36–45 from the late 1970s and early 1980s. References 46–47 provide more recent discussions. In this section, we address only the question of whether there are alternative designs which are as informative as a randomized trial. The issue is fundamental to all discussions of randomized trials.

We will assume that the purpose of a medical trial is to make a comparative statement about the efficacy of two or more treatments. Therefore, the accuracy of this statement is important. It has been argued that this particular assumption regarding a medical trial is inappropriate. Freireich [43] has argued that a comparative trial which shows major differences between two treatments is a bad trial because half the patients have received inferior treatment. Although, in a sense, this is true, we feel that any wider perspective on clinical research will encompass a desire to know the true relative merits of different treatments.

If the purpose of a trial is to compare two treatments in a prospective fashion, then randomly assigning treatments to individuals as they enter the study should avoid many potential biasses which are present in other schemes for treatment assignment. In such a setting, some sort of randomization would usually not be objectionable.

Of more concern is the trial in which a new treatment is compared to an old treatment when there is information available about the efficacy of the old treatment through historical data. Patients who receive the old treatment are the controls against whom the patients who receive the new treatment are compared. Use of the historical data for comparisons with data from the new treatment will shorten the length of the study because all patients can then be assigned to the new treatment.

The advent of statistical procedures, such as the regression models of chapter 13, which can adjust the comparison of two treatments for differing distributions of other prognostic factors in the two treatment arms, has possibly made the use of historical controls more appealing. This is because randomization appears to be unnecessary as a mechanism for ensuring comparability of the treatment arms. The weak point in this reasoning is that absolute faith is being placed in the mathematical form of the statistical model and in the choice of prognostic factors. Changes in patients and patient care from one period to another may be quite subtle, but they may generate an apparent treatment effect because the proper adjustment for such changes is unknown.

Using data from the National Wilms' Tumor Study Group, Farewell and D'Angio [42] examined these two approaches to treatment comparisons. In the first National Wilms' Tumor Study (NWTS-1), Group II and III patients were randomly assigned to three treatment groups – two single-drug chemotherapy regimens (A and B) and a double-drug regimen (C) which used both actinomycin-D and vincristine. The results of the study indicated that regimen C was the better treatment with respect to both relapse and survival. In the second National Wilms' Tumor Study (NWTS-2), regimen C was compared with a three-drug regimen (D) which added adriamycin. A total of 142 patients were entered into NWTS-1 on regimen C; from NWTS-2, data were available on 177 Group II and III patients who had received regimen D and on 179 patients who had received regimen C.

Two important prognostic factors for Wilms' tumor were histology (favorable and unfavorable) and nodal involvement. Other factors of lesser importance were age (>2 years) and tumor weight (>250 g). Farewell and D'Angio [42] chose to analyze the response variable relapse-free survival using the relative risk regression model

$$d_j(t; \underset{\sim}{x}) = d_{0j}(t)\exp(\sum b_i x_i)$$

Table 18.1. The results of two proportional hazards analyses of NWTS data: (**a**) using historical controls; (**b**) using concurrent controls

Covariate	Estimated regression coefficient	Estimated standard error	Test statistic
a Historical controls			
Regimen D	−0.21	0.29	0.72
Positive nodes	0.58	0.31	1.87
Tumor weight >250 g	−0.10	0.45	0.22
Age >2 years	0.44	0.39	1.13
b Concurrent controls			
Regimen D	−0.58	0.27	2.15
Positive nodes	0.88	0.28	3.14
Tumor weight >250 g	0.16	0.43	0.37
Age >2 years	0.27	0.37	0.73

Adapted from Farewell and D'Angio [42]. With permission from the publisher.

discussed in chapter 13. In all the analyses described below, the effect of histology was accounted for by stratification and the remaining factors were included in $\underset{\sim}{x}$, the vector of explanatory covariates.

Table 18.1 presents the results of two separate analyses. In the first of these studies, Farewell and D'Angio compared the relapse-free survival rates of the 177 patients receiving regimen D in NWTS-2 with the 142 patients receiving regimen C in NWTS-1. Their second analysis was based only on NWTS-2 data and compared the 177 patients receiving regimen D with the 179 patients receiving regimen C. The former analysis, which is based on historical controls, indicated no beneficial effect from regimen D, whereas the analysis based on concurrent controls indicated a significant, beneficial effect. More extensive modelling did not alter these conclusions. A comparison of the relapse-free survival rates of the patients who received regimen C in the two studies did not reveal any significant treatment difference ($p = 0.09$).

The final decision concerning the superiority of regimen D relative to regimen C depended on many factors involved in total patient care, for example, short- and long-term complications caused by the therapies. Historical data should not be ignored, but in the NWTS a better basis for a decision was provided by using both past and concurrent information rather than historical data alone. In fact, the design of the third National Wilms' Tumor Study reflected a conclusion that some advantage for D, measured in disease-free survival, had been established; however, there were questions about the potentially lethal, delayed cardiotoxicity of the regimen that were still unanswered. Therefore, NWTS-3 compared a modified, more intensive version of regimen C with a modified regimen D.

The results of a single trial are unlikely to resolve a controversy as involved as that of historical controls versus concurrent controls. This example does illustrate, however, some of the differences that could occur between the two types of studies.

The use of historical controls is often advocated, as in the NWTS, when a series of studies for treatment of a particular disease is being planned. The best treatment group in the most recent study becomes the control group for the succeeding study. Such an approach may be particularly prone to problems. By using the data from the best treatment in one study, the efficacy of that treatment will probably be overestimated; therefore, in a subsequent study, the results of an analysis which uses historical controls and one involving randomized controls could be different. To illustrate this principle, let us suppose that there is no real difference between two treatments, but in a particular trial one treatment, by chance, produces better results. If the same treatment is applied in a subsequent trial, it will probably generate poorer results than it did in the first trial. And even when there is a real difference between treatments, the danger of overestimating the efficacy of the better treatment is present.

Some research by Sacks et al. [45] suggests that the use of any available historical data as a control for a new treatment can lead to overestimates of the new treatment's effectiveness. In a literature review, Sacks et al. found eight randomized clinical trials and 21 trials which used historical information to compare surgical and medical treatment for coronary artery disease. Only one of the randomized clinical trials identified a significant difference in overall survival between the two treatment groups. Nearly all the trials based on historical controls found the surgical treatment to be better. Table 18.2, which is taken from Sacks et al. [45], compares long-term survival in the six randomized and nine historical trials which provided such data. The pooled historical trials show both a higher survival for surgical patients and a lower survival for medical patients. When the historical trial data are adjusted to have the same overall proportions of one-, two- and three-vessel disease as the randomized trials, the difference between the medical and surgical groups is decreased, but is still larger than the corresponding difference in the randomized trials. This adjustment was only possible for six of the historical trials.

Table 18.2. Pooled estimates of overall survival in clinical trials of medical versus surgical treatment of coronary artery disease [45][1]

	Number of studies	Number of patients	Percent survival			
			1 year	2 years	3 years	4 years
Randomized trials	6	18,861				
Surgical			92.4	89.6	87.6	85.3
Medical			93.4	89.2	83.2	79.8
Historical trials	9	9,290				
Surgical			93.0	92.2	90.9	88.3
Medical			83.8	78.2	71.1	65.5
Surgical adjusted[2]			93.7	92.5	91.2	87.4
Medical adjusted[2]			88.2	82.2	70.9	67.7

[1] Reprinted by permission of the publisher.

[2] Adjusted to have the same proportions with one-, two- and three-vessel disease as the randomized trials.

In reporting the results of their more recent investigations concerning the possible consequences of relying on non-randomized studies, Deeks et al. [47] concluded:

'Results of non-randomised studies sometimes, but not always, differ from results of randomised studies of the same intervention. Non-randomised studies may still give serious misleading results when treated and control groups appear similar in key prognostic factors. Standard methods of case-mix adjustment do not guarantee removal of bias. Residual confounding may be high even when good prognostic data are available, and in some situations adjusted results may appear more biased than unadjusted results. …

The inability of case-mix adjustment methods to compensate for selection bias and our inability to identify non-randomised studies which are free of selection bias indicate that non-randomised studies should only be undertaken when randomised controlled trials are infeasible or unethical.'

Although we have not discussed all the issues involved in opting for a randomized clinical trial, we strongly suggest that the use of historical controls can be a potentially misleading approach to treatment comparisons. Other researchers are similarly convinced. This fact, and the evident scientific advantages of concurrent, randomized controls must be given serious consideration, together with the ethical issues, in designing clinical trials.

As a final historical comment, we note that the concept of randomized experiments was proposed in order to avoid the many problems which arise in the analysis of non-randomized data. These problems were aptly described by George Yule [48], who was quoted by Irwin [49] in his presidential address to the Royal Statistical Society:

> 'The unhappy statistician has to try to disentangle the effect from the ravelled skein with which he is presented. No easy matter this, very often, and a matter demanding not merely a knowledge of method, but all the best qualities that an investigator can possess – strong common sense, caution, reasoning power and imagination. And when he has come to his conclusion the statistician must not forget his caution; he should not be dogmatic. "You can prove anything by statistics" is a common gibe. Its contrary is more nearly true – you can never prove anything by statistics. The statistician is dealing with the most complex cases of multiple causation. He may show the facts are in accordance with this hypothesis or that. But it is quite another thing to show all other possible hypotheses are excluded, and that the facts do not admit of any other interpretation than the particular one he may have in mind'.

18.5. Intention to Treat

For good clinical reasons or otherwise, some patients in a randomized study may not receive the treatment to which they have been randomized. When the trial data are analyzed, the default strategy for many, if not most, trials should be that all patients who were randomized to a particular treatment are regarded as belonging to that treatment group, whether or not they actually received the intended treatment. Thus, the randomized comparisons being made in the analysis are actually between the intent to treat with one treatment versus the intent to treat with another.

This default strategy leads to what is known as an '*intention-to-treat*' analysis. Such an approach is sometimes justified as relevant to any 'pragmatic' trial in which two interventions are being compared on the grounds that what is of practical interest is the efficacy of the attempt to intervene. Some alternative analysis strategies have been proposed and may be sensible in certain trials, especially in addition to an intent-to-treat analysis. So-called 'explanatory' trials, which focus particularly on the treatments specified in the trial protocol, and equivalence trials, which are discussed in §19.3 and for which an intention-to-treat analysis is conservative, may choose not to make an intention-to-treat analysis the primary evaluation of the study. If so, this decision should be specified in the protocol.

Patients who are randomized to treatment and subsequently found to be ineligible for the trial represent a special case, and their exclusion from a primary analysis of the trial data can often be justified.

Table 18.3. The factorial design for Stage II patients in the third National Wilms' Tumor Study

	No radiation	2,000 rads
Vincristine and actinomycin-D	Regimen W	Regimen X
Vincristine, actinomycin-D and adriamycin	Regimen Y	Regimen Z

18.6. Factorial Designs

The majority of clinical trials are designed, primarily, to answer a single question. This is often an unnecessary restriction on the design of a trial, especially for diseases which require multi-modal therapy.

For example, when the third National Wilms' Tumor Study was being designed, there were two questions of interest concerning Stage II, favorable histology, patients. One question concerned the chemotherapy comparison of two- and three-drug regimens which was mentioned in §18.4; the other was whether post-operative radiation was necessary for these patients. Since the number of cases of Wilms' tumor is small and the relapse-free survival rate is very high, two separate trials to address these questions were not feasible.

Both questions can be answered, however, in a factorial design. The schematic layout for the design is shown in table 18.3. Patients are randomized among four regimens, and the sample size of the study need not be much larger than that required to answer either question separately. The radiation question is addressed by comparing regimen W to regimen X and regimen Y to regimen Z. The chemotherapy comparison is based on W versus Y and X versus Z. With this design, it is also possible to detect a synergistic (or antagonistic) interaction between the two modalities, although if such an effect is suspected, it might be necessary to increase the sample size somewhat.

We will not describe the details of analyzing such studies; nevertheless, factorial designs pose no major analytical problems. Therefore, if the flexibility of such a trial design is attractive, researchers should not be reluctant to consider using designs of this type.

18.7. Repeated Significance Testing

In chapter 16, we discussed the problem of multiple comparisons. A related problem in the analysis of clinical trials is known as repeated significance testing.

Table 18.4. The overall probability of a significant test with repeated hypothesis testing [50]

Nominal significance level	Number of repeat tests								
	1	2	3	4	5	10	25	50	200
0.01	0.01	0.018	0.024	0.029	0.033	0.047	0.070	0.088	0.126
0.05	0.05	0.083	0.107	0.126	0.142	0.193	0.266	0.320	0.424
0.10	0.10	0.160	0.202	0.234	0.260	0.342	0.449	0.524	0.652

Reprinted by permission of the publisher.

When a clinical trial is ongoing, it is common, and ethically necessary, to prepare interim analyses of the accrued data. If one treatment can be shown to be superior, then it is necessary to stop the trial so that all patients may receive the optimal treatment. Unfortunately, the more frequently the study data are examined, the more likely it is that a 'statistically significant' result will be observed.

Table 18.4, taken from McPherson [50], illustrates this effect by showing the overall probability of observing a significant result at three nominal significance levels when a test is repeated differing numbers of times. Although this table is based on 'some fairly rigid technical assumptions' and may not be directly relevant to all clinical trials, it illustrates clearly that multiple tests at the same nominal significance level can be very misleading. For example, if we conduct ten analyses which test for a treatment difference at the nominal significance level of 0.05, the chance of falsely detecting a treatment difference is nearly 0.20, not 0.05.

There is a fair amount of statistical literature concerning formal trial designs which adjust for the effect of repeated significance testing; this area of research is known as sequential analysis. To a large extent, this predominantly theoretical work has had little effect on the actual design of medical trials. We believe that this is the case because much of the formalism does not accurately reflect the conditions under which many medical trials are conducted. A formal significance test is often one of many components in the decision to continue or stop a trial. Nevertheless, some of the more recent research in sequential analysis has greater potential for application and is affecting the design of clinical trials.

The main purpose of this section has been to make the reader aware of a frequently occurring problem in medical studies. A clinical trial should not be stopped as soon as a significant result at the 5% level has been detected. When data are constantly re-examined, and updated, the advice of a statistician should be sought before any major decisions are made on the basis of an analysis which ignores the effect of repeated statistical testing.

18.8. Sequential Analysis

The conventional view of a clinical trial can be regarded as a 'fixed sample design'. This means that a sample size is determined at the planning stage, and that the trial results are analyzed once the specified sample size has been achieved. However, as the previous section has indicated, the usual monitoring of a clinical trial often makes it 'de facto' a sequential experiment with repeated analyses over time. In this section, we give a brief introduction to some actual sequential designs. Because of technical details which we choose not to discuss, we recommend that a statistician be consulted before initiating a sequential trial. Nonetheless, we hope this section will provide useful background material for interested readers.

Most sequential designs start with the supposition that the primary comparison of the clinical trial can be represented by a test statistic. We shall represent this statistic by Z to suggest that, under the null hypothesis of no treatment difference, it is usually normally distributed with mean 0 and variance 1. For example, Z might be the usual ratio of the estimated regression coefficient associated with treatment to its estimated standard error. At any point in time during the trial, Z can be calculated.

The approach to sequential design advocated by Whitehead [51] is to consider what we might expect to see, if the null hypothesis is true, and if Z was observed or calculated continuously over time. While this is clearly impractical, it is an approach which leads to reasonable procedures that can be slightly modified to reflect the usual monitoring strategy. The essential characteristic of the design ensures that if there is no treatment difference, the overall probability, for the complete trial, of concluding that the data are not consistent with the null hypothesis is a specified significance level α. The value represented by α would often be the customary 5% level of significance. A decision that the data are inconsistent with the null hypothesis is frequently referred to as 'rejecting the null hypothesis'. Thus, in a sequential design of the type described above, the probability of rejecting the null hypothesis, on the basis of Z, sometime during the trial, is equal to α. By way of comparison, in a trial of fixed sample design a single significance test at level α is performed at the end of the trial. Since the technical details of Whitehead's approach are beyond the scope of this book, we will not discuss it further.

A second approach, known as group sequential designs, acknowledges that analyses will usually take place at specified times and presents a design based on a plan to perform a fixed number of analyses, say K, at distinct times. Group sequential designs which parallel the continuous procedures of Whitehead [51] choose a testing significance level for the jth test which is the same for all tests and such that the overall probability of rejecting the null hypothesis, if it is true, is equal to α. Thus, for example, a design which involved four planned analyses and an overall significance level of 5% would perform a significance test at each analysis at a testing significance level of 0.018.

We are sympathetic to the arguments advanced by Fleming et al. [52] that treatment differences observed in the early stages of a trial may occur for a variety of reasons, and that the primary purpose of a sequential design is to protect against unexpectedly large treatment differences. Therefore, Fleming et al. advocate using group sequential designs which preserve the sensitivity to late-occurring survival differences that a fixed sample design based on a single analysis would have. In addition, they argue that if the final analysis of a group sequential design is reached, then one would like to proceed, as much as possible, as if the preliminary analyses had not been done and a fixed sample design had been used.

To achieve these ends, Fleming et al. present designs in which the level of significance at which an intermediate analysis is performed increases as the trial progresses, and such that the testing level of significance for the final analysis is close to the overall level of α. Their proposal fulfills the ethical requirement of protecting patients while not creating substantial additional difficulties in the data analysis. The designs are characterized by K, the number of planned analyses, α, the overall significance level, and by $\mu\alpha$, the probability of terminating the trial early if the null hypothesis is true. The fraction μ is, in some sense, the proportion of the overall probability of rejecting the null hypothesis which is used up prior to the final analysis. If we denote the testing levels of significance for the K analyses by $\alpha_1, \alpha_2, ..., \alpha_K$ then specifying μ is equivalent to specifying the ratio of α_K and α, i.e., $R = \alpha_K/\alpha$. This ratio indicates how close to the overall level α the final analysis is to be performed, and reflects the effect which the sequential nature of the design is allowed to have.

Table 18.5 presents a subset of the designs described in Fleming et al. [52]. The table covers the cases specified by $\alpha = 0.05$, K = 2, 3, 4, and 5 and $\mu = 0.1$, 0.3 and 0.5. For example, if three analyses were planned and it was important to keep the ratio R high, i.e., $\mu = 0.1$ so that $R = 0.04831/0.05 = 0.97$, then the testing significance levels would be $\alpha_1 = 0.00250$, $\alpha_2 = 0.00296$ and $\alpha_3 = 0.04831$. On the other hand, if a more liberal stopping criterion was desirable, the design with $\mu = 0.5$ would result in testing significance levels of $\alpha_1 = 0.01250$, $\alpha_2 = 0.01606$ and $\alpha_3 = 0.03558$ with $R = 0.03558/0.05 = 0.71$.

Table 18.5. Testing significance levels $\alpha_1, \ldots, \alpha_5$ for some group sequential designs

K	μ	α_1	α_2	α_3	α_4	α_5
2	0.1	0.00500	0.04806	–	–	–
	0.3	0.01500	0.04177	–	–	–
	0.5	0.02500	0.03355	–	–	–
3	0.1	0.00250	0.00296	0.04831	–	–
	0.3	0.00750	0.00936	0.04292	–	–
	0.5	0.01250	0.01606	0.03558	–	–
4	0.1	0.00167	0.00194	0.00233	0.04838	–
	0.3	0.00500	0.00612	0.00753	0.04342	–
	0.5	0.00833	0.01047	0.01306	0.03660	–
5	0.1	0.00128	0.00144	0.00164	0.00219	0.04806
	0.3	0.00379	0.00447	0.00527	0.00642	0.04319
	0.5	0.00634	0.00776	0.00931	0.01134	0.03691

The overall level of significance is 0.05, and 0.05 μ is the probability of terminating the trial early, if the null hypothesis is true, at any of the K analyses.

Adapted from Fleming et al. [52]; it appears here with the kind permission of the publisher.

Table 18.6. Anticipated results from the use of a sequential design proposed by Fleming et al. [52] for a clinical trial of extensive stage small-cell lung cancer

Date	Total number of patients randomized to		Number of deaths observed	Testing significance level for early termination	Log-rank p-value observed
	regimen A	regimen B			
9/12/77	19	17	15	0.007	0.013
5/05/78	30	32	30	0.008	0.214
11/12/78	32	33	45	0.010	0.701
7/15/79	32	33	60	0.040	0.785

Adapted from Fleming et al. [52]; it appears here with the kind permission of the publisher.

Table 18.6 is abstracted from Fleming et al. and reports the results of a clinical trial of extensive stage small-cell lung cancer. Two chemotherapy regimens, denoted by A and B, were to be compared. Calculations based on a fixed sample design to compare death rates using the log-rank test suggest that the study would require about 60 deaths. The nature of these calculations is outlined in chapter 17. If we assume that K = 4 log-rank analyses are planned during the trial, one every 15 deaths, and if we also require $R = \alpha_4/\alpha = 0.8$, then the four testing significance levels which result are $\alpha_1 = 0.007$, $\alpha_2 = 0.008$, $\alpha_3 = 0.010$ and $\alpha_4 = 0.040$.

From table 18.6 it can be seen that although there was a relatively large treatment difference early in the trial, this difference would not have been sufficient to stop the trial. Moreover, by the end of the trial no treatment difference was apparent.

This section is not intended to be a comprehensive treatment of the topic of sequential designs. Additional study of the subject, and consultation with a statistician, would be essential before embarking on a clinical trial which involves a sequential design. However, we do feel that the designs proposed by Fleming et al., which we have described, are consistent with the usual practice of clinical trials. Therefore, they may be of interest to some readers.

19

Further Comments Regarding Clinical Trials

19.1. Introduction

Chapter 18 provides a basic introduction to clinical trials. While we hope that the discussion there was realistic, it necessarily adopted a fairly simple view of a clinical trial. In this chapter, we raise several issues that are somewhat more complex, but which we feel it worthwhile to bring to the reader's attention.

19.2. Surrogate Endpoints

For diseases that involve a lengthy delay between the initiation of treatment and the determination of its outcome, keeping the trial as short as possible is often an important goal. One way to achieve this is to base the analysis of the trial on a 'surrogate endpoint' or 'surrogate marker' which is thought to be an early indicator of outcome. For example, in trials designed to evaluate the ability of an experimental drug to delay the progression of HIV disease, investigators might consider using the level of CD4+ lymphocytes or a measure of viral load as a surrogate for the 'harder' clinical endpoints such as the onset of AIDS or death. The underlying logic is that if decreased levels of CD4+ cells are associated with increased risk of an AIDS diagnosis, then a consistently depressed CD4+ count could be regarded as a relevant clinical endpoint on which to base analyses. In drug regulation, there are obvious advantages to patients and pharmaceutical companies if biomarkers can be used as valid surrogates of the ultimate clinical benefit of treatments and thus shorten the time span of the trial. Likewise, the use of surrogate endpoints may be considered when the response of primary interest is difficult to observe, expensive to measure, or involves a dangerous invasive procedure.

Prentice [53] defines a valid surrogate endpoint as 'a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.' This definition is very useful. For example, it identifies that a surrogate endpoint for one clinical trial may not be useful for another clinical trial involving the same primary endpoint, but different treatments. Nevertheless, the definition is a fairly restrictive one that is rarely satisfied in practice.

The dangers involved in using even surrogate endpoints that appear to be sensible have been highlighted by Fleming [54], who discussed two trials. Ventricular arrhythmias are a risk factor for subsequent sudden death in individuals who have had a recent myocardial infarction. The drugs encainide and flecainide were widely used in treatment because their antiarrhythmic properties had already been established. Nevertheless, the Cardiac Arrhythmia Suppression Trial was initiated and, based on 2,000 randomized patients, established that the death rate associated with using these drugs was nearly three times the corresponding rate for placebo controls.

Fleming's second example is that of a trial concerning the role of γ-interferon in the treatment of chronic granulomatous disease (CGD). Phagocytes from CGD patients ingest microorganisms normally but fail to kill them due to an inability to generate a respiratory burst that depends on the production of superoxide and other toxic oxygen metabolites. The disease results in a risk of recurrent serious infections. Since γ-interferon was thought to be a macrophage-activating factor that could restore superoxide anion production and bacterial killing by phagocytes in CGD patients, a trial was designed initially that would involve one month of treatment and would use as response variables endpoints that depended on the two surrogate markers superoxide production and bacterial killing. Ultimately, the investigators decided to administer γ-interferon for a year in a trial that measured the occurrence of serious infections as the outcome. This trial established the effectiveness of γ-interferon in reducing the rate of serious infection. However, when the biological surrogate marker data that had been proposed as outcome variables in the initial design were analyzed, treatment with γ-interferon had no apparent effect. Thus, a trial based on these two surrogate markers as endpoints would have failed to detect an effective treatment.

These two examples cited by Fleming, demonstrating the potential for both false positive and false negative results in conjunction with the use of surrogate endpoints, provide a useful caution against the uncritical use of outcomes other than those of primary clinical interest. Careful evaluation of surrogate endpoints should precede their use in a clinical trial; Burzykowski et al. [55] describe methods for such an assessment. Note, in particular, that a strong correlation between the surrogate endpoint and a true endpoint does not ensure the surrogate will be good.

There is some potential, perhaps, in the joint use of surrogate markers and primary outcome variables but initial investigations into appropriate methodology have not been as promising as had been hoped. We encourage readers who are considering the use of surrogate endpoints to keep the following in mind: the question asked is the one that gets answered!

19.3. Active Control or Equivalence Trials

When an effective standard therapy exists, a new experimental treatment may be investigated because of reduced toxicity, lower cost or some other characteristic that would make it, the experimental therapy, the treatment of choice *if* its efficacy was equivalent to or better than the standard. The design of this type of trial – an active control or equivalence trial – cannot be based on the usual significance test because, as is the case for all significance tests, failure to reject a null hypothesis of no treatment difference in a clinical trial does not establish the equivalence of the treatments thus compared. A large significance level associated with a test of this null hypothesis indicates that the data gathered during the trial are consistent with the null hypothesis. However, in order to learn what size of treatment effects remain plausible in light of the data, it is necessary to look at confidence intervals.

Fleming [56] advocates the use of confidence intervals to analyze active control trials. First, he identifies a point, denoted by e, that represents overall therapeutic equivalence of the treatments under study. This point is defined in terms of an efficacy outcome, but its value will depend on other considerations such as toxicity and cost. Next, Fleming specifies a quantity that represents the departure from e that would lead to the conclusion that one treatment under study was superior to the other; this quantity is represented by the Greek letter δ. Finally, Fleming defines the relative efficacy of a placebo to the standard therapy to be ϕ; the value of ϕ is assumed to be known from previous studies.

Although specifying e, δ, and ϕ clearly involves considerable subjectivity, these three characteristics together provide a framework for analyzing an equivalence trial. The precise nature of this structure is illustrated in figure 19.1. The horizontal axis in the figure indicates the relative efficacy of the experimental treatment to the active control and could represent a mean difference, an odds ratio, a relative risk, or any other appropriate, comparative measure. For example, if the lower limit of a 95% confidence interval for the relative

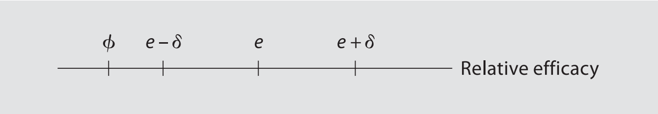


Fig. 19.1. Key aspects of the framework proposed by Fleming [54] for analyzing an equivalence trial.

efficacy of the experimental treatment to the active control exceeds $e - \delta$, and it also exceeds ϕ, then the experimental treatment will be considered equivalent or superior to the active control. In this case, the evidence indicates that the experimental treatment is not inferior to the standard and provides some benefit compared to no treatment whatsoever, i.e., placebo control. Similarly, if the upper limit of a 95% confidence interval for the relative efficacy of the experimental treatment to the active control is less than $e + \delta$, the hypothesis that the experimental therapy is superior will be rejected. If the lower limit exceeds $e - \delta$ and the upper limit is less than $e + \delta$, the two treatments are considered equivalent. In many situations there may be no particular interest in equivalence per se, and it is simply that the lower limit of the confidence interval exceeds $e - \delta$ that is of interest. In this case, the terminology *non-inferiority trial* is sometimes used instead of equivalence trial.

Note that the logic underlying sample size calculations for equivalence trials is somewhat different from that outlined in chapter 17, and a statistician or specialized references should be consulted. Essentially, the sample size has to be sufficient to ensure that the width of the confidence interval for the relative efficacy is small enough to allow values below $e - \delta$ to be excluded with high probability if the relative efficacy is at least e.

Adopting a sequential design can often improve an equivalence clinical trial. As well, it may be desirable to incorporate other outcomes, for example toxicity, into the formal analysis of the trial. Methodology to facilitate these extensions has been developed. Since the intent of this section was solely to introduce the different sort of logic that must be considered when questions of equivalence are entertained, we will not attempt to discuss sequential methods for equivalence trials.

19.4. Other Designs

In our discussion of the design of clinical trials so far, we have assumed the most common situation, which is sometimes termed a *parallel group design*. In such trials, patients are individually randomized to two or more treatment groups. Two other designs with which readers might come into contact are *cross-over trials* and *cluster randomized studies*.

Cross-over designs can be used when the medical condition of interest is expected to be ongoing so that patients can be treated sequentially with more than one treatment. For example, different symptomatic treatments for asthma could be made available to a patient during successive months, and the effectiveness of symptom relief could be recorded for each time period. If we assume that two treaments are under study, say A and B, and that two time periods will be used for the trial, then the classic two-period cross-over design would enroll patients into the trial and then randomly assign them to receive treatments in the order AB or BA.

The potential advantage of a cross-over design is that treatments can be compared within patients rather than between patients, as a parallel group design dictates. As we previously indicated in §9.4.1, the use of paired data will usually lead to more precise comparisons. The analysis of cross-over designs in general is a specialised topic, and we will not attempt to provide details here. A key aspect of the design and corresponding analysis is that in addition to estimating a measure of treatment effect, allowance can be made for different levels of response in the two 'periods' of the design. However, if the effect of one or both treatments can 'carry-over' from the first period of the design to the second, then a cross-over design is not recommended.

Extensions to the two-period cross-over format are possible. In particular, increasing the number of treatment periods offers increasing flexibility in the design and analysis of such trials. A very special case of extending the cross-over design is the so-called *n-of-1 trial* in which only a single patient is involved but the number of treatment periods is large enough to allow one to estimate the treatment effect for that individual patient. The practical use of n-of-1 trials is likely to be limited; however, a multi-period cross-over design can be regarded as a series of n-of-1 trials.

Cluster randomized trials, which are also called group randomized trials, arise when the randomization of individual patients is not possible. For example, varying the treatment provided from patient to patient within the same medical center or practice might be logistically difficult. To study the effectiveness of educational materials designed to discourage cigarette smoking, the same educational program might have to be given to all students in a class or even to all students in a school. In such cases, a group of study subjects are effectively randomized together to the same intervention and may even receive the intervention at the same time and in the same place.

To analyze cluster randomized trials, treatments must be compared at the level where randomization occurred, e.g., between medical centers, between classes or schools. The essential feature of the data that requires attention in the analysis is the correlation between the observed outcomes for study subjects who were randomized together. A common approach to the analysis of such trials is to use random effects models; these were discussed briefly in §14.3 as one way of analyzing correlated longitudinal data. We do not propose to discuss the details here, but note that an appropriate analysis of such trials will provide less precise estimates of any treatment effects than one could obtain from an individually randomized trial involving a similar total number of subjects. Thus, it is important not to ignore the clustering.

A slightly more subtle type of clustering occurs when a trial is individually randomized, but the subjects are nevertheless clustered. A simple example of this phenomenon would be a multi-centre trial in which the study subjects treated at the same centre form a cluster of potentially correlated observations. For such trials it is generally recognized that some adjustment for the effect of each centre is required; often, this adjustment is achieved through stratification.

A potentially more complex example corresponds to situations that involve clustering by the health professional who is delivering a treatment. In a trial of two treatments, A and B, it could be that the delivery of treatment A requires special training, and therefore separate sets of health professionals will treat patients in the two arms of the trial. Alternatively, it may be that health professionals are able to treat patients in both arms of the trial, but there are still relatively few treatment providers compared to the number of patients. Random effects models provide an approach to the analysis of such trials that adjusts appropriately for the clustering associated with each health professional.

While we have not provided details of any statistical methods for trials with clustering, we hope that readers have been made aware of the situations in which clustering can arise. In such situations, statistical advice should be sought for trial planning, including appropriate sample size calculations, and subsequent data analysis.

19.5. Multiple Outcomes

The existence of a single primary outcome that adequately characterizes response to treatment considerably simplifies the problem of designing a suitable study. However, to insist on a single outcome variable in every clinical trial is much too narrow a view. If a particular therapy is to be used in a comparative clinical trial but sufficient toxicity information is not available, then collecting both efficacy and toxicity data during the course of the study would be of interest. In a stroke prevention trial there may be interest in determining, for various stroke types, the treatment benefit that can be characterized in terms of both severity and location. Similarly, in arthritis studies disease activity is reflected in various measured variables including pain, active joint counts, strength, and mobility. It is sensible, therefore, to consider how multiple outcome variables affect the design and analysis of clinical trials.

One approach to the formal analysis of multiple outcomes is to adopt a single summary response that combines the individual responses in a clinically sensible manner. For example, this approach is often used in quality-of-life studies when responses are combined across various dimensions. In many cases, however, the summary index that is adopted is based more on mathematical convenience than on clinical relevance. If so, then to attach a clinical interpretation to a treatment effect that is measured in terms of the summary variable can be very difficult. For example, in a cancer trial with two outcomes, tumor response and survival time, a moderate significance level, say $p = 0.03$, could result from a dramatic treatment effect on one of the outcome variables, and a limited effect on the other. The same significance level could also be due to two consistently moderate treatment effects, one for each outcome variable. In the former case, subsequent interest may focus on the particular response associated with the large treatment effect. Therefore, the summary index is of limited value since separate analyses of tumor response and patient survival must be completed.

It follows, therefore, that the analysis of a number of separate response variables is of more interest than a simple analysis of a summary index. In designing a clinical trial, the framework that we previously described in chapter 17 is frequently adopted. This methodology includes the notion of a Type I error rate, which can be interpreted as the chance of obtaining a false-positive conclusion from the trial. If testing is carried out at a particular significance level, α, and if the null hypothesis is true, then the probability of declaring that a significant result has been observed is α. If a number of significance tests, relating to a variety of outcome variables, are performed, all with a false-positive error rate of α, then the proportion of trials involving *one or more* false-positive errors overall may be much more than α.

There are a variety of statistical methods for controlling the overall error rate, which is defined as the rate of making one or more false-positive errors, in experiments that involve multiple hypothesis tests. We introduced this topic during a discussion of HLA and disease relationships in §16.6. While the procedures that we previously described seem appropriate in that context, and in others, they are not necessarily suited to a clinical trial with multiple outcomes.

Cox [57] has pointed out that probabilities pertaining to the simultaneous correctness of many statements may not always be directly relevant, particularly when a specific response is of interest. This suggests that in designing a clinical trial that formally involves two or more responses, multiplicity adjustments may not be necessary if the results of individual hypothesis tests are interpreted separately and have consequences that involve different aspects of the treatment prescription. This basically means that the treatment effect for each specific response is of interest. If hypothesis tests are directed primarily at inferences concerning several different responses, individually, it is reasonable to specify a maximum tolerable error rate for each specific hypothesis test.

The following example illustrates what we mean in a concrete situation. Pocock et al. [58] reported a colorectal cancer trial in which two systemic chemotherapy regimens were evaluated, based on the two outcome measures tumor response after two months of treatment and survival time. Analysis of a 2×2 table concerning the association between tumor response and treatment generated a test statistic with a value of 4.50. If the null hypothesis of no treatment effect is true, this statistic should have a χ_1^2 distribution; therefore, the resulting significance level of the test was 0.034. A log-rank test comparing the survival experience of the patients treated with the two regimens generated an observed value for the test statistic of 2.11; the corresponding significance level was 0.15. To combine these results concerning both outcome variables, Pocock et al. [58] calculated a global test statistic that is related to one of the approaches to multiple comparisons that we discussed in §16.6. The significance level of the combined test was 0.038. Alternatively, if Pocock et al. had chosen to use Bonferroni multiplicity corrections in order to jointly analyze the treatment effect for both tumor response and survival, then both hypothesis tests would have yielded significance levels in excess of 5%, and no evidence for a treatment benefit would normally be claimed.

It may be doubtful, however, that the clinical importance of this colorectal cancer trial could be determined solely on the basis of this global test of significance. A trial that provides very strong evidence of a treatment benefit in terms of tumor response and corresponding weak evidence of benefit in terms of survival would be quite different from a study that exhibited weak evidence of a benefit with respect to tumor response and strong evidence in terms of survival. Thus, the two responses seem to be of interest in their own right, and the effect of treatment on both tumor response and survival time would be relevant to patient care.

Similarly, an analysis that is based on Bonferroni-type adjustments seems inappropriate because it protects against one or more false-positive results. This approach would only be relevant if a treatment was to be recommended for use provided one or more of the separate significance tests was significant. The relevance of such a conceptual framework seems questionable if the results of separate hypothesis tests can have individual implications for different response variables.

A major factor leading to the increased number of trials with multiple outcomes has been the formal use of quality-of-life measures and health economic measures in determining the value of treatments. Since these measures often involve many different outcomes, the number of responses that must be incorporated at the design stage of a trial can be quite large. One possible strategy in this case is to recognize that these indices are effectively grouped according to various characteristics, e.g., quality-of-life dimensions, and to construct summary measures or global statistics within the groups thus defined. The preceding discussion in this section is most relevant to the joint consideration of such summary measures or global statistics. It is also relevant to the joint consideration of, say, quality-of-life measures with traditional clinical outcomes such as survival. Previous attempts to specify relevant summary measures such as quality-adjusted life years [see Torrance, 59] have met with some criticism [see Cox et al., 60], and the use of global test statistics in this context seems generally inappropriate.

We do not intend readers to infer from the preceding discussion that we support the examination of endless outcome variables at the end of a trial in order to find at least one that produces a statistically significant result. Instead, we hope that this brief discussion of multiple outcomes has provided some guidance for researchers involved in the design of trials when separate clinical concerns are reflected in a small number of outcome variables.

19.6. Multiple Treatment Arms

While the simplicity of the two-treatment, randomized trial is attractive, the complexity of modern therapies means that, increasingly, clinical trials with multiple treatment arms are being designed. These arms might involve various dosages of an experimental treatment, cumulative combination therapies, or simply multiple different treatments. Such designs are often viewed as an efficient method of simultaneously assessing the efficacies of a variety of treatment options. For example, several active treatments can be compared to one another, and/or to a common control group receiving standard medical care.

Consider a trial involving K treatment arms. In principle, an investigator may want to compare all pairs of treatments. However, when formal significance tests are involved, the basic rule to remember is that only K – 1 significance tests can be performed on such data unless the multiple-comparisons problem is satisfactorily addressed.

Clearly, situations do arise in medical studies when there is understandable vagueness about the precise comparison or comparisons that might be of greatest interest. In fact, in §16.6 we briefly described a preliminary epidemiologic study to investigate the relationship between HLA alleles and a specific disease in which all pairwise comparisons might indeed be of interest. However, it is harder to imagine that such non-specific hypotheses could correspond to the reality of a comparative clinical trial. Treatment arms are usually introduced into a trial for one or more particular reasons.

One approach to the analysis of multi-armed trials would be to predicate all subsequent analyses of treatment differences on the outcome of a global test for the absence of any difference between the treatment groups. This corresponds to the method that we previously outlined in §16.6. An alternative strategy is to specify the contrasts of interest at the design stage of the trial. This ensures that no treatment comparison is based on observed differences between treatment groups that actually may be spurious. In this case, each comparison is examined separately, and any concern about the overall false-positive error rate corresponding to these multiple tests would be minimal.

The reasons for introducing each of the various treatment arms in a clinical trial usually will define a small number of clinical questions of primary interest. If the number of questions is not excessive, then each question can be considered separately. Since global tests of the null hypothesis of no treatment difference tend to be rather unfocussed, it seems very reasonable, in comparative clinical trials, to proceed directly to the consideration of these various separate questions.

In the analysis of variance literature, considerable attention is devoted to the topic of orthogonal contrasts. Mathematically speaking, this amounts to posing, and subsequently testing, separate questions that are, in some sense, independent of each other. However, in the clinical trial setting, the use of orthogonal contrasts and the associated advantage of independence are unattractive for practical reasons, since this independence is often achieved at the expense of adopting hypotheses that have less clinical relevance. A pragmatic approach to this issue is to limit the number of contrasts considered to one less than the number of treatment arms. If the questions thus posed are not independent in the mathematical sense, then the interpretation of each depends on the other questions addressed in the same way that, in a regression model, a test of the hypothesis that the regression coefficient for a particular covariate is zero depends on the information conveyed by the other variables included in the regression model. Thus, the effect represented by a particular contrast is examined after adjusting for the possible effects represented by the other contrasts. However, if the number of treatment comparisons exceeds $K - 1$ when the total number of treatment arms is K, then it is no longer clear that each comparison is providing separate information. In this case, the view that each significance test can be interpreted separately to assess the treatment information contained in the trial is questionable.

Finally, readers should note that a concern regarding *any* false-positive conclusions may be of considerable importance if the trial results are directly linked to decisions. For example, some so-called Phase II clinical trials are initiated in order to select a few treatments, or perhaps only one, for further investigation. In such cases, the overall false-positive error rate is very important and a formal multiple-comparisons method should be used to analyze the trial results. This type of study is essentially a 'selection' procedure. In many trials, however, the motivation for the different arms of the design will not be selection, per se, but rather an examination of different aspects of treatment strategies. It is in these circumstances that a moderate number of clinically relevant hypotheses can be specified, and that multiple testing procedures need not be adopted.

19.7. Sequential Designs for Efficacy-Toxicity Trials

The subject of efficacy-toxicity trials represents a natural application of the notion that more than one outcome may be of interest in a particular clinical trial. However, this section undoubtedly involves the most technical treatment of any topic in the book. Readers who found the discussion of sequential analysis in the previous chapter heavy going are strongly advised to omit this section and proceed to the next section.

Due to the urgent nature of some comparative clinical trials, therapies may sometimes be evaluated before well-established toxicity information is available. This situation occurs even in therapeutic contexts where new treatments tend to be increasingly toxic. In such trials, efficacy and toxicity data must be collected simultaneously. Since ethical and practical reasons dictate that both efficacy and toxicity responses must be monitored, these endpoints are often considered equally important. This situation represents a particular example of clinical trials with more than one outcome of interest. For example, the evaluation of therapeutic interventions for the treatment of HIV-related non-Hodgkin's lymphoma is a setting in which efficacy and toxicity responses are a primary concern for clinical researchers. The standard chemotherapy regimens that are effective in reducing mortality due to lymphoma further compromise the already-suppressed immune system in HIV-seropositive patients. Consequently, aggressive chemotherapy is often accompanied by an increased incidence of opportunistic infections.

In §19.5 we indicated that researchers often want to maintain the ability to examine two or more outcomes separately. In this section, we describe a method of simultaneously monitoring efficacy and toxicity responses that was developed by Cook and Farewell [61]. Their work extends the sequential methods that we discussed in the previous chapter to include trials that involve multiple outcomes. Unfortunately, the problem of designing such a trial is sufficiently complex that we cannot provide a complete explanation of the procedure described in [61]. In addition, the statistical literature contains alternative approaches to designing efficacy-toxicity trials that we will not attempt to discuss here. However, we hope the following discussion gives readers some sense of the issues that arise when more than one outcome is of interest in a clinical trial. Clinicians who become involved in designing trials of this nature should definitely consult a statistician.

We introduce θ_1 to represent the relative efficacy of the new treatment compared to the control, and θ_2 to denote the corresponding relative toxicity. Recall that we introduced the notion of relative efficacy in §19.3 and indicated there that relative efficacy could be a mean difference, an odds ratio, a relative risk, or any other appropriate, comparative measure. Relative toxicity is defined similarly for adverse events. For clarity of exposition, we will assume that each relative measure is a mean difference; then $\theta_1 = \theta_2 = 0$ means that the two therapies are equally effective and equally toxic. We suppose that the investigators are interested in testing jointly the hypotheses H_{10}: $\theta_1 = 0$ versus H_{1a}: $\theta_1 \neq 0$ and H_{20}: $\theta_2 \leq 0$ versus H_{2a}: $\theta_2 > 0$. The first hypothesis test specifies that the experimental therapy has the same efficacy as the control, whereas the second hypothesis test indicates that the experimental treatment is no more toxic than the control method. These tests will be based on the observed values of appropriate test statistics pertaining to θ_1 and θ_2 whenever a planned analysis of the accumulated trial data is carried out. We also suppose that the researchers are willing to stop the trial if H_{10} is rejected, or H_{20}, or both hypotheses. That is, the trial will stop early for one or more of the following reasons: (i) increased efficacy of the experimental treatment, i.e., $\theta_1 > 0$, (ii) reduced efficacy of the experimental treatment, i.e., $\theta_1 < 0$, or (iii) excess toxicity of the experimental treatment, i.e., $\theta_2 > 0$. The trial will not stop early if evidence of reduced relative toxicity, i.e., $\theta_2 < 0$, is observed.

Notice that the hypotheses concerning relative efficacy and toxicity are not the same. Because evidence of reduced relative toxicity does not constitute sufficient grounds to end the trial early, the hypothesis test concerning θ_2 involves alternatives that lie in one direction only, i.e., $\theta_2 > 0$. The resulting test is commonly called a one-sided significance test concerning θ_2. We will not discuss such tests in detail, but note that there are two different reasons why such tests arise. The first pertains to situations where alternatives to the null hypothesis in one direction are known to be impossible. For example, in a study of a recombination fraction in genetics, the difference between the recombination fraction and the usual value of 1/2 may be the primary quantity of interest. However, it is known that the fraction cannot be less than 1/2; hence, alternative values of the difference that are negative are impossible.

The second reason, which corresponds to the case of θ_2, the relative toxicity parameter, arises when alternative values of the parameter of interest may lie on either side of the boundary between the null and alternative hypotheses but only values in one direction are deemed to be interesting. Thus, the experimental therapy may be more toxic, or less so, than the control treatment. However, evidence of reduced relative toxicity does not constitute sufficient grounds to stop the trial early.

Inevitably, the premature termination of a clinical trial affects the accumulation of data concerning the outcomes of interest. In a trial involving both efficacy and toxicity, few researchers would dispute that early stopping is appropriate when the experimental therapy is less effective than the control treatment. In this case, collecting additional information regarding the relative toxicity of the experimental treatment is unnecessary. Similarly, if the experimental therapy is sufficiently toxic, the study would be halted for ethical reasons, thereby eliminating the need to collect any more information concerning efficacy. On the other hand, if the trial was halted prematurely because of the superior efficacy of one of the therapies, additional toxicity data would be collected and analyzed in subsequent corroborative or postmarketing surveillance trials. Thus, the primary reason for monitoring the toxicity response is to allow for the early termination of trials that involve extremely toxic experimental therapies. However, if a trial is terminated early because of favourable efficacy results, it may be the case that insufficient data were collected to enable researchers to properly assess the relative toxicity. Depending on the frequency and severity of the toxicity outcome, collecting sufficient data could be important to establish the safety of the experimental treatment. This possibility does not represent a major problem if the group sequential procedures for bivariate response data developed by Cook and Farewell are regarded as stopping *guidelines* [62].

In chapter 18 we discussed sequential designs for a single response of interest that were first proposed by Fleming et al. [52]. Those designs were characterized by the fraction μ. If the null hypothesis that the control and experimental therapies are equally effective is true, μ represents the proportion of the overall probability of rejecting the null hypothesis that is used up prior to the final analysis of the trial data. If α denotes this overall probability, i.e., the probability that the trial results in a false-positive conclusion, then the value

$\mu\alpha$ represents the probability of terminating the trial early, when the null hypothesis is true.

Although we did not indicate it in chapter 18, these same sequential designs can also be characterized in terms of the probability of stopping the trial at each planned analysis because of a false-positive result. If the trial design involves at most K analyses of the observed data, then we can denote the probability of stopping the trial at the kth consecutive analysis, if the null hypothesis is true, by π_k. This would mean that, at the kth analysis of the data accumulated during the trial, π_k is the probability that the observed results will lead to the false-positive conclusion that one treatment is superior. The designs that we described in chapter 18 are based on equating π_k to the fraction $\mu\alpha/(K - 1)$, for the initial $K - 1$ planned analyses, and setting $\pi_K = 1 - \mu\alpha$. Thus, there is an equal probability of stopping the trial because of a false-positive conclusion at each analysis prior to the final one. Although this particular choice of values for $\pi_1, \pi_2, ..., \pi_K$ is not strictly necessary, it simplifies the problem of designing and describing a sequential clinical trial that involves only one response, and is generally used.

Likewise, in a sequential trial involving both efficacy and toxicity responses, we can define various probabilities that pertain to stopping the trial because of a false-positive result when the two treatments being compared are equivalent with respect to efficacy and the experimental therapy is no more toxic than the control, i.e., when the relative efficacy, θ_1, is equal to 0 and the relative toxicity, θ_2, is at most 0. Let π_{1k} represent the probability of stopping the trial at analysis k solely because of a false-positive result with respect to efficacy. Similarly, π_{2k} represents the probability of stopping at analysis k solely because of a false-positive result concerning the toxicity response, and π_{12k} represents the probability of stopping at analysis k because of false-positive results with respect to both the efficacy and toxicity analyses. Then π_k, the overall probability of stopping at analysis k because of a false-positive result on either the efficacy or the toxicity response, is the sum of the three probabilities π_{1k}, π_{2k}, and π_{12k}, i.e., $\pi_k = \pi_{1k} + \pi_{2k} + \pi_{12k}$.

In order to determine the monitoring protocol for a sequential trial involving both efficacy and toxicity, we need to identify the values of the probabilities π_{1k}, π_{2k}, and π_{12k} at each of the K planned analyses, as well as the overall significance levels pertaining to testing of each of these two important outcomes. Moreover, this assignment of values must be such that the entire protocol has an overall significance level of α. The first step in this process is to partition α, the overall significance level, among the K planned analysis stages. Equivalently, we must specify values of $\pi_1, \pi_2, ..., \pi_K$ such that $\pi_1 + \pi_2 + \ldots + \pi_K = \alpha$.

Next, it is necessary to further specify the probability of stopping the trial at each analysis stage because of a false-positive result with respect to either, or both, of the efficacy and toxicity outcomes. A convenient way of doing this is to determine, for each planned analysis, the probability that a false-positive result with respect to efficacy is the reason for stopping the trial, *if* the trial is terminated prematurely at that particular planned analysis. Since the probability of stopping the trial at analysis stage k is π_k and the probability of stopping and falsely rejecting H_{10}, the hypothesis concerning relative efficacy, is $\pi_{1k} + \pi_{12k}$, then the requirement can be achieved by assigning values to the ratios $f_1, f_2, \ldots, f_K$, where

$$f_k = \frac{\pi_{1k} + \pi_{12k}}{\pi_k}.$$

Thus, choosing the stopping probabilities, $\pi_1, \pi_2, \ldots, \pi_K$, determines the overall significance level of the entire protocol, while specifying the ratios $f_1, f_2, \ldots, f_K$ determines how this probability is allocated to the separate analyses of efficacy and toxicity at each planned evaluation of the observed data. Notice that this method of identifying the probabilities π_{1k}, π_{2k}, and π_{12k} treats the null hypotheses concerning efficacy and toxicity somewhat differently. The overall significance level of the efficacy analysis is only determined indirectly, through the specification of the ratios $f_1, f_2, \ldots, f_K$, via the equation

$$\alpha_1 = \sum_{i=1}^{K} \{\pi_{1i} + \pi_{12i}\}.$$

The corresponding overall significance level for the toxicity analysis, if the null hypothesis is true, is equal to

$$\alpha_2 = \sum_{i=1}^{K} \{\pi_{2i} + \pi_{12i}\}.$$

Although we cannot demonstrate it here, the value of α_2 that results depends on the correlation between the efficacy and toxicity responses.

The following three examples illustrate the consequences, for the trial protocol, of possible choices for the ratios $f_1, f_2, \ldots, f_K$. Suppose the trial design specifies at most five analyses, i.e., $K = 5$. By assigning $f_1 = 1 = f_2 = f_3 = f_4 = f_5$ we reduce the bivariate stopping rule to a univariate sequential design; setting $f_k = 1$ eliminates the analysis of the toxicity response at stage k. On the other hand, the values $f_1 = 0.1$, $f_2 = 0.3$, $f_3 = 0.5$, $f_4 = 0.7$, $f_5 = 0.9$ correspond to analyses early in the trial that are more sensitive to effects concerning the toxicity response; the later analyses place more emphasis on effects related to the efficacy outcome. Finally, the values $f_1 = 0$, $f_2 = 1$, $f_3 = 0$, $f_4 = 1$, and $f_5 = 0$ correspond to a group sequential procedure that directs analyses to each of the two outcomes in an alternating pattern.

Another issue that an investigator must face when designing a group sequential trial is the problem of specifying the appropriate overall significance level for the entire protocol. This question is directly related to our previous discussion in §19.5 concerning multiple outcomes. In the example that follows, we will use a sequential trial design that involves an overall significance level of 0.05; however, readers should recognize that, when the trial explicitly involves two outcomes about which separate conclusions will be drawn, specifying an overall significance level of 0.05 may be inappropriate. For example, a design that closely approximates the choice $\alpha_1 = 0.05$ and $\alpha_2 = 0.025$, values that probably would be used in separate studies of the same efficacy and toxicity endpoints, might be preferable. Although we have not chosen to describe them here, there are alternative approaches to the sequential analysis of multiple endpoints that will explicitly determine α_1 and α_2. Any of these methods may be particularly useful in specific situations; see, for example, Cook [63].

To provide a concrete illustration of some of the features of efficacy-toxicity trial design, we conclude this section with an example based on a study of kidney transplantation. The purpose of the trial was to compare the usual cyclosporin treatment that is used to prevent rejection of the transplanted organ with an experimental therapy that combined cyclosporin with a ten-day regimen of Campath. Graft survival was the efficacy response of interest, but there was an additional concern regarding toxicity that would manifest itself through the occurrence of infection. The test statistics that were to be used to analyze the efficacy and toxicity responses were log-rank test statistics comparing the time to graft failure – the efficacy response – in the two treatment groups and the time to infection – the toxicity response – as well. Data concerning the occurrence of these events have been extracted from summaries of the trial that were provided by the investigators, but for the purposes of illustration, entry times have had to be assigned since these were no longer available.

Consider a hypothetical efficacy-toxicity design involving three possible analyses that might have been developed for this kidney transplant study so that $K = 3$, and fix an overall significance level for testing both the efficacy and the toxicity hypotheses at 5%. Scheduled analyses will occur 250, 500 and 750 days after the start of the trial. When the null hypotheses that the two therapies are equally effective and that the experimental therapy is no more toxic than the control treatment are true, the values of the stopping probabilities at each planned analysis will be $\pi_1 = 0.01$, $\pi_2 = 0.01$, and $\pi_3 = 0.03$. Because of the particular concern about toxicity, the two interim analyses will be directed primarily at detecting excessive toxicity by setting $f_1 = 0.1$ and $f_2 = 0.4$; choosing $f_3 = 0.9$ redirects the focus of the final analysis to the efficacy response. Such a sequential design will yield overall significance levels for testing the null hypotheses concerning efficacy and toxicity of $\alpha_1 = 0.032$ and $\alpha_2 = 0.018$, respectively.

The resulting significance levels for testing the separate efficacy and toxicity outcomes, using the accumulated information concerning graft failure and the occurrence of infection 250 days after the start of the trial, are 0.001 and 0.009, respectively, i.e., $\pi_{11} + \pi_{121} = 0.001$ and $\pi_{21} + \pi_{121} = 0.009$. Figure 19.2 shows the Kaplan-Meier curves, by treatment group, for the efficacy and toxicity outcomes. The associated log-rank statistics that compare the experience of the two treatment groups with respect to graft failure and the occurrence of infection 250 days into the trial are 38.7 and 13.7. The nominal significance levels associated with these observed values of the log-rank statistic are 5×10^{-10} and 0.0002, respectively. Therefore, on the basis of these results from the first 250 days of follow-up in the kidney transplant study, both the null hypothesis that the treatments are equally effective in preventing transplant failure and the null hypothesis that the cyclosporin/Campath regimen is no more toxic than cyclosporin alone would be rejected; the study would be terminated prematurely at the first analysis. As figure 19.2 indicates, after just 250 days of experimentation the investigators would be able to conclude that the experimental therapy involves a reduced risk of graft rejection; however, the simultaneous risk of infection is distinctly greater when the use of cyclosporin is combined with a ten-day regimen of Campath.

19.8. Stochastic Curtailment

The situation may arise when early termination of a clinical trial is considered for reasons other than established efficacy of a treatment or unacceptable toxicity. A common cause is poor recruitment, and this reason can be more compelling if there is little evidence of treatment differences in the data already collected.

In such a situation, it is sometimes argued that the information which has already been gathered should be used to consider the likely outcome of the trial if it were continued in order to achieve the planned sample size for the study. The trial will usually have been designed to achieve a specified level of power, e.g., 80 or 90%, to detect a treatment difference of interest. The treatment difference usually will be represented by a parameter of interest, say θ, in some statistical model. The null hypothesis of no treatment difference will correspond to $\theta = 0$ and the treatment difference of interest, which is often termed the smallest clinically relevant difference, will correspond to a specific nonzero value for θ, say $\theta = \delta$. Within this framework, the available data can be

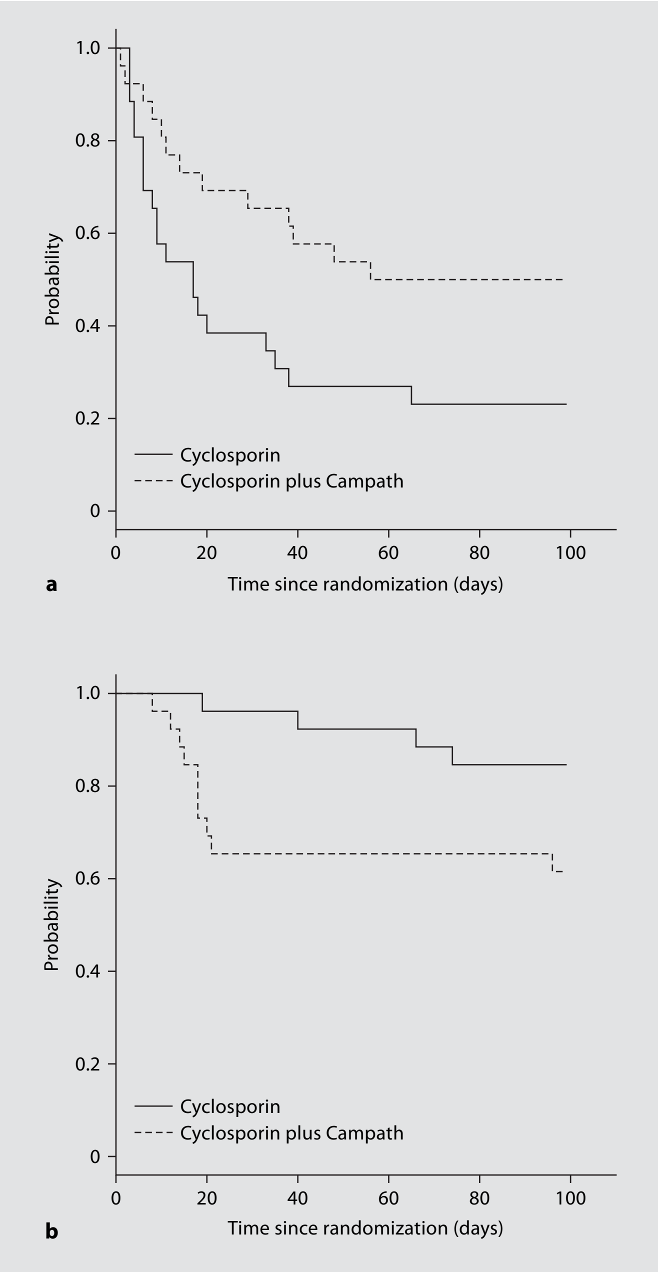


Fig. 19.2. Estimated Kaplan-Meier curves for the efficacy-toxicity trial of cyclosporin versus cyclosporin plus Campath in kidney transplantation patients. **a** Graft failure. **b** Occurrence of infection.

used to calculate the probability that, if the trial recruits to its planned sample size, the null hypothesis regarding θ will be rejected at a given level of statistical significance, provided the true treatment difference is δ. This predicted value is known as conditional power, i.e., the probability of rejecting the null hypothesis in favour of the better treatment, for the given value of $\theta = \delta$, conditional on the currently available data. It is common, and most sensible, to restrict attention to rejection of the null hypothesis in favour of the better treatment. However, considering the rejection of the null hypothesis in favour of either treatment usually does not change the conditional power very much.

Note that calculating the conditional power is not simply a case of re-calculating the power with the currently estimated value of θ. Such a calculation has no particular justification or merit.

Suppose a clinical trial has been undertaken to assess pain relief as measured on a visual analogue scale. The smallest clinically relevant difference, in the context of the intervention being studied, is set at a difference in means of 10 mm. If the trial investigators assume that a common standard deviation in the two patient groups is 25, and specify a Type 1 error rate of 5%, a total of 133 subjects in each group is required to achieve a power of 90%, i.e., the probability of rejecting the null hypothesis in favour of the better treatment, if the difference between the means of the two treatment groups in the trial is indeed 10 mm, is 90%. Near the end of the planned recruitment period only 90 subjects per group have been recruited, representing 68% of the desired total. Information from these 180 subjects can be used to estimate the probability of rejecting the null hypothesis in favour of the better treatment if recruitment were to continue until the full 266 subjects had been enrolled, and the treatment difference was in fact 10 mm. If this estimated value is low, then it follows that there is little point in continuing the trial. Indeed, the complement of conditional power – 100% minus the conditional power – is sometimes called the futility index.

For example, at the stage when 90 subjects per group have been recruited, if the observed mean pain scores are 37.0 and 35.0 in favour of the intervention arm, with a standard deviation of 25 in each group, then around 68% of the required information is available, corresponding in this case to having 68% of the required sample size. If the assumed treatment difference, θ, is equal to 10 mm, the conditional power is less than 19%. Even if the true difference is 10 mm, if the trial is continued it is unlikely that the results of any final analysis of the accumulated data will lead to rejection of the null hypothesis of no treatment effect.

Some individuals have claimed that clinical trials can be monitored on the basis of conditional power; however, we believe there are a number of points to consider.

(a) The use of conditional power places central importance on the *predicted* result of a final significance test, which remains uncertain. It does not focus on what is known from the currently available data. For example, if on the basis of current data the value $\theta = \delta$ is excluded from an appropriately calculated confidence interval then that knowledge is more consistent with the type of information that a trial is designed to achieve and may provide a more compelling argument for termination.

(b) If we do not have information in the available data that is clinically useful, then continuing the trial will provide a more precise estimate of any possible treatment difference, and this knowledge could be valuable for a variety of purposes. Thus the term futility index which is associated with conditional power may be misleading. Although the trial may be futile with respect to achieving a significant result, providing clinically useful information is a worthy goal.

(c) If clinical trials are routinely stopped on the basis of conditional power, this change in practice may bias any meta-analysis in which they are subsequently included. Chapter 20 introduces the concept of a meta-analysis. The bias could arise if trials are included without suitably adjusting for the stopping criterion.

(d) Low conditional power alone provides no ethical reason for stopping a trial if useful information can be achieved by continuation, i.e., there is no treatment difference that provides an ethical reason for stopping.

These four points lead us to recommend caution in the use of conditional power for monitoring clinical trials. However, subject to the usual caveats about the simplistic assumptions that underlie power calculations, the estimated value of conditional power does provide valid information. Thus, it might well be appropriate to 'take comfort' from conditional power calculations if, for other reasons such as poor recruitment, it seems essential to stop a trial early.

20

Meta-Analysis

20.1. Introduction

In previous chapters, we have discussed various issues in the design and analysis of clinical trials. However, our attention has focused on individual trials. In recent years there has been a surge of interest in methods to combine the information from a number of trials, or studies, that all provide information on the same question of interest.

For many years in clinical and epidemiological research, this activity was largely qualitative and was communicated through review articles. In these, 'experts' would survey the literature and attempt to draw general conclusions. The authors of such review articles rarely attempted to provide a numerical basis for their conclusions. Authors and readers alike recognized that inevitably such reviews would be selective in the studies they reported, perhaps most obviously through the usual de facto inclusion of only published research.

The attempt to formalize, and improve, this process of synthesizing research has led to the widespread use of what are often termed 'systematic reviews'. A very important part of any systematic review is the specification of methods to identify, and characterize the quality of, data for inclusion in the review. Search procedures to identify all relevant studies, published and unpublished, are initiated along with methods of data acquisition. Any reader who wishes to undertake a systematic review should consult the appropriate literature to understand this aspect of the process.

In this chapter, we have the limited aim of highlighting some of the statistical issues that arise in the analysis of data acquired as part of a systematic review. The statistical method is called meta-analysis. From our perspective, meta-analysis is one component of a systematic review, although this same term is sometimes used more generally to refer to a systematic review.

20.2. Background

Meta-analysis involves the analysis of data concerning the same question of interest that has been extracted from multiple primary research studies. Three major aims of a meta-analysis are:

(1) To use quantitative methods to explore heterogeneity of study results;

(2) To estimate overall measures of associations or effects;

(3) To study the sensitivity of conclusions to the method of analysis, publication bias, study quality or other features of interest.

Normally, a meta-analysis will be based on a single outcome measure that can be extracted from all the studies identified. There are procedures that sometimes can be used if trials have different outcomes but all measure the efficacy of the same intervention. One suggested approach in this situation involves combining standardized 'effect sizes' or perhaps p-values. However, such meta-analyses are often problematic and are best avoided if possible.

In other cases, all trials may use the same outcome measure but the only information available from the published reports consists of data summaries, for example estimated treatment effects and their standard errors. Then the typical meta-analytic procedures will combine the various estimates, weighting their influence on the overall estimate of a treatment effect according to their precision, as measured by the corresponding standard errors.

The most useful situation is to have original data from each study in order to undertake a combined analysis. This allows the data to be analyzed jointly in the most efficient manner and gives scope for more comprehensive analyses. This is the situation that we will use for illustration in §20.4, but the qualitative issues that arise will be the same for any approach that is adopted.

A meta-analysis may be based on any of the types of data – including binary, continuous, time-to-event, count, ordinal or nominal measurements – that we have discussed in previous chapters. In addition to identifying the type of data, investigators who are planning a meta-analysis will need to choose a suitable outcome measure. Some possibilities are summarized in table 20.1; not all of these have been mentioned in previous chapters.

A comprehensive discussion of analyses for these various possibilities is well beyond the intended scope of this chapter. However, in §4 we will focus on methods that might be used for binary data when the outcome measure for the effect of interest is an odds ratio. The general approach that we adopt is similar for other types of data and effect measures, but for details the reader will need to consult a statistician or another source.

Before presenting formal analysis methods for binary data however, we offer some general comments on study heterogeneity.

Table 20.1. Possible outcome measures for various types of data

Type of data	Possible outcome measures			
Binary	Odds ratio	Relative risk	Risk difference	Number needed to treat
Count	Expected value	Relative rate		
Continuous	Original scale	Transformed scale	Standardized effect measure	
Ordinal or nominal	Specialized methods			

20.3. Study Heterogeneity

Even if the primary question of interest is the same, there will undoubtedly be plenty of variation between studies with respect to patient populations, medical care systems and similar factors. However, any reference to study heterogeneity in the context of a meta-analysis usually pertains to the more important issue of heterogeneity in the effect of interest. Thus, as part of a meta-analysis, statistical tests for the presence of differences in effects from study to study will usually be needed.

Such tests of heterogeneity should be undertaken in the early stages of a meta-analysis. Simplistically, these will often take the form of a formal significance test. We will use the simple dichotomy of such a test in order to discuss some of the pertinent issues, but presume that readers, with statistical assistance, will also undertake a more comprehensive review of all the data collected.

Assume first that no heterogeneity from one study to another is detected in the sense that the significance level of a suitable formal test does not fall below a chosen threshold such as 5%. In this case it is essential to remember that lack of statistical significance does not imply lack of heterogeneity but rather only lack of evidence of heterogeneity. Since the power of most tests of heterogeneity can be quite low, this reminder implies that a subsequent blind acceptance that effects across studies are homogeneous is probably unwise. In addition, readers should be aware that more sensitive tests of heterogeneity can generally be identified if specific hypotheses about the possible source of any heterogeneity are prescribed. For example, rather than testing for any possible differences between studies, one could choose to compare the effects for randomized and unrandomized trials, if both types of studies are included in the meta-analysis. As with any other method of data analysis however, for such a test to be a valid method of comparison the hypotheses must be specified in advance and not one suggested by a previous examination of the data.

If evidence of heterogeneity is detected, the immediate next step would be an examination of the nature of that heterogeneity. Distinguishing between qualitative and quantitative heterogeneity can be useful. The former term refers to an effect that is present in some studies but absent in others, or perhaps even displaying opposing directions. On the other hand, quantitative heterogeneity refers to less dramatic numerical variation in the size of the effect across studies. Once any heterogeneity that is present has been carefully characterized, the impact that it may exert on any analyses needs to be identified. For example, is the concept of a suitably defined, single estimated treatment effect still reasonable?

20.4. An Illustrative Example

For illustration purposes, consider the meta-analysis of studies with binary outcome data where Y = 1 corresponds to some kind of failure and Y = 0 denotes a success. We assume the question of interest involves a comparison of two groups that can be distinguished by a binary explanatory variable X; X = 1 indicates a treated subject and X = 0 corresponds to a control subject.

An example of such data is given in table 20.2, which is extracted from figure 6 in Peto et al. [64]. The table summarizes the results of 11 randomized clinical trials of prolonged antiplatelet therapy. The primary outcome of interest is death, and each trial has a treated and control arm. The size of the different studies varies markedly, and the overall death rates fluctuate as well.

As the discussion in chapter 11 indicated, logistic regression is a natural method for analyzing binary data such as the investigators in these 11 randomized trials have collected. For the simple situation we have described, such a model corresponds to the equation

$$\Pr(Y = 1 \mid x) = \frac{\exp(a + bx)}{1 + \exp(a + bx)},$$

where b is the log odds ratio and exp(b) is the corresponding odds ratio that characterizes the effect of the antiplatelet therapy.

In this fairly simple probability model, the parameter a corresponds to the overall rate of death. However, this rate will likely differ from study to study, and so corresponds to the heterogeneity between studies which is not linked to the possible benefit of treatment that we discussed in the previous section. This situation strongly suggests that we should stratify any analysis by study.

Table 20.2. Summary data from 11 randomized trials of prolonged antiplatelet therapy

Trial	Treatment		Control	
	number of deaths	total number randomized	number of deaths	total number randomized
1	58	615	76	624
2	129	847	185	878
3	244	1,620	77	406
4	154	1,563	218	1,565
5	395	2,267	427	2,257
6	88	758	110	771
7	39	317	49	309
8	102	813	130	816
9	38	365	57	362
10	65	672	106	668
11	9	40	19	40

Adapted from Peto et al. [64]. It appears here with the kind permission of the publisher.

An appropriate stratified logistic regression model that is similar to equation (11.2) would correspond to the equation

$$\Pr(Y=1\,|\,x)=\frac{\exp(a_i+bx)}{1+\exp(a_i+bx)}$$

where the subscript i indexes the 11 studies. This version of the logistic regression model assumes that the overall death rate can vary arbitrarily from study to study; however, the effect of the treatment, which is measured by b, is the same in all studies. Such a stratified logistic regression model is foundational for a meta-analysis because allowing for variation in the pattern of outcomes from study to study is essential in order to avoid the type of problem that we identified in §11.4. Such study heterogeneity is seldom of particular interest itself.

If we adopt this model to analyze the data summarized in table 20.2, the estimated value of b is –0.28, with a standard error of 0.04. The corresponding estimated odds ratio is exp(–0.28) = 0.75. Since the magnitude of the estimate is seven times larger than its standard error, a test of the null hypothesis b = 0, i.e., antiplatelet therapy has no effect on the death rate, is highly significant. The 95% confidence interval for b is (–0.37, –0.20), and the corresponding interval for the odds ratio associated with treatment is (0.69, 0.82). All the values in this interval represent a substantial treatment benefit.

This foundational logistic regression model is sometimes called a 'fixed effect model' because b, the effect of interest, is assumed to be constant, or fixed, across all 11 studies. We can relax this assumption by adopting the model

$$\Pr(Y = 1|x) = \frac{\exp(a_i + b_i x)}{1 + \exp(a_i + b_i x)},$$

where the treatment effect parameter b is allowed to vary from study to study. This added flexibility is achieved by adding a subscript i on b as well as on the parameter a.

Such a model has two primary uses. The first involves using the model as a means of testing for heterogeneity of the treatment effect from study to study. This action corresponds to testing the null hypothesis that all the b_is are the same, represented by the hypothesis $b_i = b_0$ for all i; the value b_0 represents a common, unknown constant. Typically a test of this type would be calibrated against a χ^2 distribution with degrees of freedom equal to one less than the total number of studies. For a logistic regression model, this hypothesis can be investigated using a likelihood ratio test similar to the one mentioned in §16.6. Based on the data summarized in table 20.2, such a likelihood ratio test statistic has an observed value of 14.18 and should be compared with a χ^2_{10} distribution. The resulting significance level is 0.17. Thus, the data do not provide any evidence for heterogeneity of the treatment effect between studies. We will not go into any greater detail here, since the issues concerning heterogeneity were discussed in the previous section.

The second use of this modified logistic regression model is based on assuming that the treatment effect, b_i, is random. We previously introduced this random effects assumption in §14.3, although in that instance it was the intercept term, a_i, that we assumed was random. In this case, it is the b_is that are random and come from a common distribution; the resulting model is another type of 'random effects model'. The normal distribution is a frequent choice for that common distribution of some random effect; in this case, we might suppose that the associated, unknown mean and variance are b and σ^2, respectively. Then the primary result stemming from the meta-analysis is an estimate of b, the mean of this common distribution for the random treatment effects. This estimated mean is then regarded as the 'overall' estimate of the treatment effect from the meta-analysis.

There has been considerable debate amongst statisticians as to whether a fixed or random effects model is more appropriate for a meta-analysis. Essentially, a fixed effect analysis will produce an estimate of the overall treatment effect for all studies by weighting the information from each study on the basis of its total size. The random effects analysis explicitly introduces study-to-study variability; the estimated overall treatment effect is then some sort of mean parameter. The fixed and random effects overall estimates can differ because the random effects model will allocate a larger weight to smaller studies than would be justified solely on the basis of their sample size. Perhaps the more important difference between these two approaches is that the extra variability in the random effects model will be reflected in any confidence interval for the overall estimate. Therefore, such an interval for the overall estimated outcome measure from a random effects model will be wider, and hence more conservative.

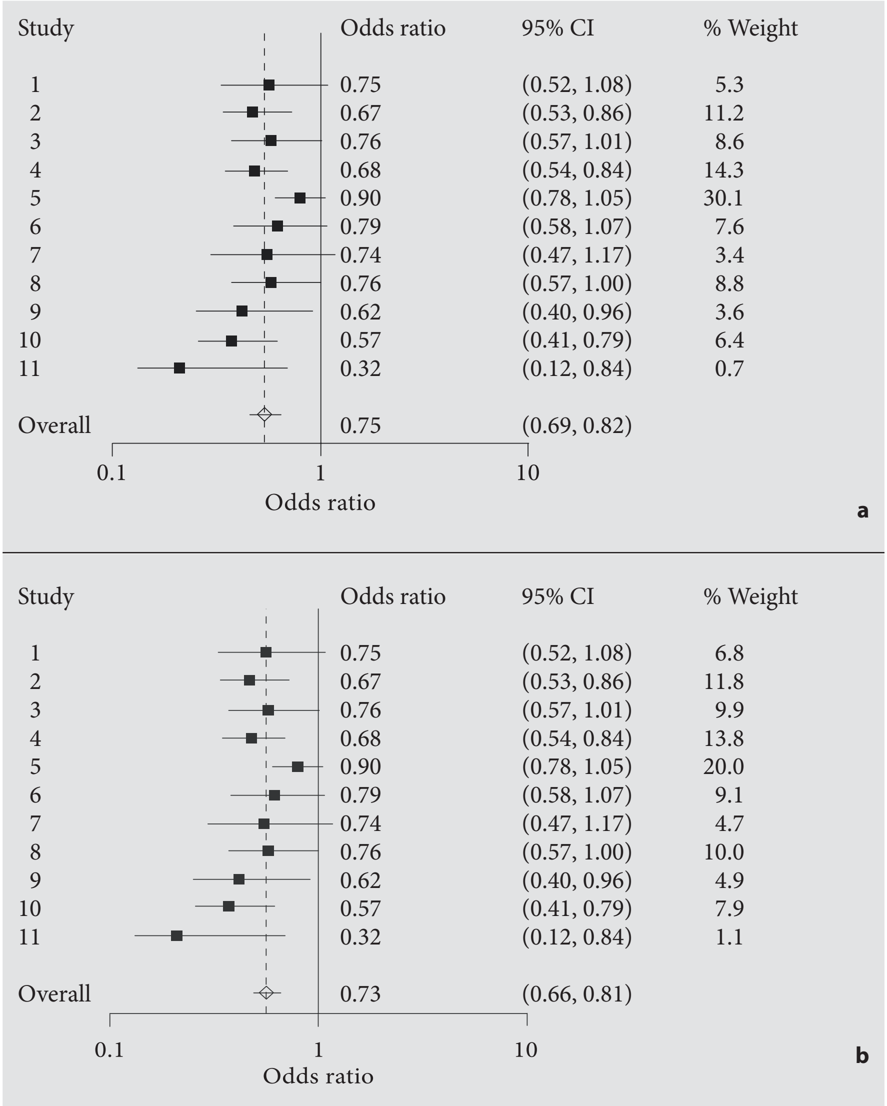
For the data in table 20.2, the overall treatment effect estimated via a random effects model corresponds to an odds ratio of 0.73 with associated 95% confidence interval (0.66, 0.81). In this case, the interval estimate is only slightly larger than its fixed effects model counterpart, and the two estimates of the odds ratio are almost identical.

We will not express a preference for one method or the other in general, but hope that we have at least indicated the nature of the difference between these two approaches to meta-analysis. If any reader is forced to make a choice between them, perhaps our discussion will prove helpful. In many situations, if there is doubt, then undertaking both analyses to determine the practical consequences of such a choice is be wise.

20.5. Graphical Displays

Graphical displays are the most common method of presenting the results of a meta-analysis. A typical figure would include effect estimates and corresponding confidence intervals from all studies incorporated in the meta-analysis, as well as the overall estimate of the effect measure and associated confidence interval.

Figure 20.1 presents two such displays, called caterpillar plots, for the fixed and random effects meta-analyses of the 11 studies summarized in table 20.2. These plots show estimated odds ratios and 95% confidence intervals for each individual study, as well as the overall estimate and associated 95% confidence interval. The first caterpillar plot is based on the results of a fixed effect analysis, and the second plot is derived from the random effects model that we discussed in the previous section. The only difference between the results from the two analyses occurs in the assigned weights and therefore also in the overall estimated odds ratio and its corresponding 95% confidence interval.



Study	Odds ratio	95% CI	% Weight
1	0.75	(0.52, 1.08)	5.3
2	0.67	(0.53, 0.86)	11.2
3	0.76	(0.57, 1.01)	8.6
4	0.68	(0.54, 0.84)	14.3
5	0.90	(0.78, 1.05)	30.1
6	0.79	(0.58, 1.07)	7.6
7	0.74	(0.47, 1.17)	3.4
8	0.76	(0.57, 1.00)	8.8
9	0.62	(0.40, 0.96)	3.6
10	0.57	(0.41, 0.79)	6.4
11	0.32	(0.12, 0.84)	0.7
Overall	0.75	(0.69, 0.82)	

Study	Odds ratio	95% CI	% Weight
1	0.75	(0.52, 1.08)	6.8
2	0.67	(0.53, 0.86)	11.8
3	0.76	(0.57, 1.01)	9.9
4	0.68	(0.54, 0.84)	13.8
5	0.90	(0.78, 1.05)	20.0
6	0.79	(0.58, 1.07)	9.1
7	0.74	(0.47, 1.17)	4.7
8	0.76	(0.57, 1.00)	10.0
9	0.62	(0.40, 0.96)	4.9
10	0.57	(0.41, 0.79)	7.9
11	0.32	(0.12, 0.84)	1.1
Overall	0.73	(0.66, 0.81)	

Fig. 20.1. Caterpillar plots of the meta-analysis of 11 randomized trials of prolonged antiplatelet therapy. **a** Fixed effects analysis; **b** random effects analysis.

The estimated odds ratios and confidence intervals from the individual studies are listed beside each caterpillar plot, along with the weight assigned to each individual study in the calculation of the overall estimate. These weights make clear the difference between the two analyses. For example, study 5 is assigned a weight of 30% in the fixed effect analysis, but only 20% in the random effects analysis. The smallest trial, study 11, was assigned 0.7 and 1.1% weights in these two analyses, respectively.

20.6. Sensitivity

Our final comment concerning meta-analyses is that they typically involve a number of 'choices', and the robustness of any final conclusions to the decisions taken deserves to be investigated. The most critical of these choices probably relates to selecting individual studies to include in the meta-analysis, and the actual method of analysis to use. Carrying out a variety of analyses by varying these choices is certainly the simplest way to investigate the confidence that one can attach to any preferred meta-analysis.

21

Epidemiological Applications

21.1. Introduction

Lilienfeld and Lilienfeld [65] begin their text 'Foundations of Epidemiology' with the statement: 'Epidemiology is concerned with the patterns of disease occurrence in human populations and the factors that influence these patterns.' In previous chapters we have concentrated on clinical data related primarily to disease outcome. Here we attempt to provide a brief introduction to the study of disease incidence, and illustrate how some of the statistical methods which we have discussed can be applied in various epidemiological studies.

21.2. Epidemiological Studies

The cohort study is the simplest approach to the study of disease incidence. A cohort, which is a random sample of a large population, is monitored for a fixed period of time in order to observe disease incidence. Important characteristics of each individual cohort member are ascertained at the start of the study and during the period of follow-up. This information is used to identify the explanatory variables or 'risk factors' which are related to disease incidence. Thus, risk factors are measured prospectively, that is, before the occurrence of disease. The cohort study also provides a direct estimate of the rate of disease incidence in the population subgroups which are defined by the explanatory variables. Section 21.3 describes, in considerable detail, the analysis of a cohort study which uses the method of proportional hazards regression that was introduced in chapter 13.

A special case of cohort data is population-based, cause-specific incidence and mortality data which are routinely collected for surveillance purposes in many parts of the world. These data are usually stratified by year, age at death (or incidence) and sex and have been used to study geographic and temporal variations in disease. In certain instances, aggregate data on explanatory variables such as fat consumption or smoking habits may also be available for incorporation into the analysis of these vital data.

For many diseases, the collection of cohort data is both time-consuming and expensive. Therefore, one of the most important study designs in epidemiology is the case-control study. This design involves the selection of a random sample of incident cases of the study disease in a defined population during a specified case accession period. Corresponding comparison individuals (the controls) are randomly selected from those members of the same population, or a specified subset of it, who are disease-free during the case accession period. Information on the values of explanatory variables during the time period prior to case or control ascertainment is obtained at the time of ascertainment. These retrospective data are usually subject to more error in measurement than the prospective data of the cohort study; however, a case-control study can be completed in a much shorter period of time. The case-control design facilitates comparisons of disease rates in different subsets of the study population but, since the number of cases and controls sampled is fixed by the design, it cannot provide an estimate of the actual disease rates. An important variation in case-control designs involves the degree of matching of cases to controls, particularly with respect to primary time variables such as subject age. In §21.4 we present an example of a case-control study in which logistic regression is used to analyze the data.

21.3. Relative Risk Models

In chapter 13, we introduced proportional hazards regression as a method for modelling a death rate function. This same method can also be used to model the rate of disease incidence. If we denote the disease incidence rate at time t by r(t), then we can rewrite equation (13.1) as

$$\log\{r(t; \underset{\sim}{x})\} = \log\{r_0(t)\} + \sum_{i=1}^{k} b_i x_i, \tag{21.1}$$

where $\underset{\sim}{x} = \{x_1, x_2, ..., x_k\}$ refers to a set of explanatory variables to be related to the incidence rate. The function $r_0(t)$ represents the disease incidence rate at time t for an individual whose explanatory variables are all equal to zero.

To illustrate how the regression model (21.1) can be used in epidemiology, consider the cohort study reported by Prentice et al. [66]. This cohort study was based on nearly 20,000 residents of Hiroshima and Nagasaki, identified by population census, and actively followed by the Radiation Effects Research Foundation from 1958. Information concerning systolic and diastolic blood pressure, as well as a number of other cardiovascular disease risk factors, was obtained during biennial examinations. During the period 1958–1974, 16,711 subjects were examined at least once. In total, 108 incident cases of cerebral hemorrhage, 469 incident cases of cerebral infarction and 218 incident cases of coronary heart disease were observed during follow-up. The determination of the relative importance of systolic and diastolic blood pressure as risk indicators for these three major cardiovascular disease categories was a primary objective of the analysis of these data. The time variable t was defined to be the examination cycle (i.e., t = 1 in 1958–1960, t = 2 in 1960–1962, etc.). The use of this discrete time scale introduces some minor technical issues which need not concern us here.

Two generalizations of model (21.1) which were introduced in chapter 13 are of particular importance to the application of the model to cohort studies. Frequently, there are key variables for which the principal comparison in the analysis must be adjusted. For example, in the study of blood pressure and cardiovascular disease it is important to adjust for the explanatory variables age and sex. A very general adjustment is possible through the stratified version of Cox's regression model, for which the defining equation is

$$\log\{r_j(t; \underset{\sim}{x})\} = \log\{r_{j0}(t)\} + \sum_{i=1}^{k} b_i x_i, \quad j = 1, \ldots, J. \tag{21.2}$$

Equation (21.2) specifies models for J separate strata, each of which has an unspecified 'baseline' incidence rate $r_{j0}(t)$, where j indexes the strata. As in the lymphoma example discussed in §13.3, the regression coefficients are assumed to be the same in all strata. Prentice et al. [66] used 32 strata defined on the basis of sex and 16 five-year age categories.

The second generalization of model (21.1) involves the use of time-dependent covariates. In this case the equation for the analysis model is

$$\log\{r_j[t; \underset{\sim}{x}(t)]\} = \log\{r_{j0}(t)\} + \sum_{i=1}^{k} b_i x_i(t), \quad j = 1, \ldots, J. \tag{21.3}$$

In their analysis of the cohort data, Prentice et al. used blood pressure measurements at examinations undertaken before time t as covariates. Thus, for example, table 21.1 presents regression coefficients for a covariate vector, $\underset{\sim}{X}(t) = \{S(t–1), D(t–1)\}$, which represents the systolic and diastolic blood pressure measurements, respectively, in examination cycle t – 1. The table includes the

Table 21.1. The results of a relative risk regression analysis of cardiovascular disease incidence in relation to previous examination cycle systolic and diastolic blood pressure measurements; the analyses stratify on age and sex

Regression variable	Cerebral hemorrhage	Cerebral infarction	Coronary heart disease
$S(t-1)$	0.0058[a] (0.30)[b]	0.0177 (<0.0001)	0.0115 (0.003)
$D(t-1)$	0.0548 (<0.0001)	0.0046 (0.36)	–0.0046 (0.56)
Cases	92	406	187

[a] The estimated regression coefficients, $\hat{b}$, are maximum partial likelihood estimates.

[b] The values in parentheses are significance levels for testing the hypothesis $b = 0$.

Adapted from Prentice et al. [66]. It appears here with the permission of the publisher.

corresponding significance levels for testing the hypothesis $b_i = 0$. Separate analyses are presented for each cardiovascular disease classification. The incidence of other disease classifications is regarded as censoring in each analysis. This is consistent with the assumption that the overall risk of cardiovascular disease is the sum of the three separate risks.

In table 21.1, the diastolic blood pressure during the previous examination cycle is the important disease risk predictor for cerebral hemorrhage, whereas the corresponding systolic blood pressure is the more important predictor for cerebral infarction and for coronary heart disease. From table 21.1, the estimated risk of cerebral hemorrhage for an individual with a diastolic blood pressure of 100 is

$$\exp\{0.0548(100-80)\} = 2.99$$

times that of an individual whose diastolic pressure is 80. Relative risks comparing any two individuals can be calculated in a similar way for each of the disease classifications.

The assumption that the blood pressure readings during the previous examination cycle are indeed the important predictors can be tested by defining new regression vectors $\underset{\sim}{X}(t) = \{D(t-1), D(t-2), D(t-3)\}$ for the cerebral hemorrhage analysis and $\underset{\sim}{X}(t) = \{S(t-1), S(t-2), S(t-3)\}$ for cerebral infarction and coronary heart disease. For a subject to contribute to these analyses at an incidence time t, all three previous biennial examinations must have been at-

Table 21.2. The results of a relative risk regression analysis of cardiovascular disease incidence in relation to blood pressure measurements from the three preceding examination cycles; the analyses stratify on age and sex

Regression variable	Cerebral hemorrhage	Cerebral infarction	Coronary heart disease
S(t – 1)	–	0.0113 (0.001)	–0.0101 (0.06)
D(t – 1)	0.0323[a] (0.01)[b]	–	–
S(t – 2)	–	0.008 (0.03)	0.0146 (0.007)
D(t – 2)	–0.0107 (0.45)	–	–
S(t – 3)	–	0.0035 (0.30)	0.0064 (0.22)
D(t – 3)	0.0477 (<0.0001)	–	–
Cases	48	207	97

[a] The estimated regression coefficients, $\hat{b}$, are maximum partial likelihood estimates.

[b] The values in parentheses are significance levels for testing the hypothesis b = 0.

Adapted from Prentice et al. [66]. It appears here with the permission of the publisher.

tended; therefore, the analyses presented in table 21.2 involve fewer cases than the results reported in table 21.1. The more extensive analysis presented in table 21.2 suggests that the most recent systolic blood pressure measurement, S(t – 1), is strongly associated with the risk of cerebral infarction, while the next most recent, S(t – 2), shows a much weaker association. On the other hand, after adjusting for the levels of systolic blood pressure in the two preceding cycles, a recent elevated systolic blood pressure measurement is negatively associated, although marginally, with the risk of coronary heart disease. One possible explanation for this result would be the suggestion that hypertensive medication achieves blood pressure control without a corresponding reduction in coronary heart disease risk. The analysis for cerebral hemorrhage indicates that both elevated diastolic blood pressure and the duration of elevation are strongly related to the incidence of cerebral hemorrhage.

Tables 21.1 and 21.2 illustrate a complex application of the proportional hazards regression model in the analysis of cohort data. By discussing this example, we have tried to illustrate some of the possibilities for analysis which use of this model facilitates. There are a number of issues which arise in the analysis of cohort data via relative risk regression models which are beyond the scope of this book. In our view, however, these models represent a natural choice for the analysis of epidemiological cohort studies. We hope it is also clear that careful thought must be given to the form of a regression model that is used to analyze epidemiological data, and to the interpretation of the results.

21.4. Odds Ratio Models

In this section we consider the case-control study reported by Weiss et al. [67]. This study identified and interviewed 322 cases of endometrial cancer occurring among white women in western Washington between January 1975 and April 1976. A random sample of 288 white women in the same age range were interviewed as controls. The interviews were used to obtain information on prior hormone use, particularly postmenopausal estrogen use, and on known risk factors for endometrial cancer.

Let Y be a binary variable which distinguishes cases (Y = 1) from controls (Y = 0). Then Y corresponds to the event of endometrial cancer incidence during the study period. If we also define a binary explanatory variable X which is equal to 0 unless a woman has used post-menopausal estrogens for more than one year (X = 1), then we require a statistical model which relates Y to X. A natural choice in this instance is the binary logistic regression model which was introduced in chapter 11. In terms of Y and X, the defining equation for this model is

$$\log\left\{\frac{\Pr(Y=1\,|\,x)}{1-\Pr(Y=1\,|\,x)}\right\}=a+bx. \qquad (21.4)$$

As we indicated in §11.3, e^b is the odds ratio which compares the odds of being a case for a woman using estrogen to the same odds for a non-user. The probability of cancer incidence for a non-user is equal to exp(a)/{1 + exp(a)}.

In §21.2 we noted that, because the proportions of cases and controls are fixed by the study design, it is impossible to estimate the probability of cancer incidence, which depends on the parameter a, from a case-control study. Nevertheless, if case-control data are analyzed using a model like equation (21.4), then although the estimate of a has no practical value, the estimation of odds ratio parameters such as b can proceed in the usual fashion and provides valid

Table 21.3. The results of an odds ratio regression analysis of endometrial cancer incidence in relation to exposure to exogenous estrogens and other factors; the analysis stratifies on baseline age

Risk factor	Regression variable	Definition	$\hat{b}$	SE[a]	Significance level[b]
Estrogen use	X_1	1 if duration of use between 1 and 8 years; 0 otherwise	1.37	0.24	<0.0001
	X_2	1 if duration of use 8 years or greater; 0 otherwise	2.60	0.25	<0.0001
Obesity	X_3	1 if weight greater than 160 lbs; 0 otherwise	0.50	0.25	0.04
Hypertension	X_4	1 if history of high blood pressure; 0 otherwise	0.42	0.21	0.05
Parity	X_5	1 if number of children 2 or greater; 0 otherwise	0.81	0.21	0.0001

[a] Estimated standard error of $\hat{b}$.

[b] Estimated significance level for testing the hypothesis b = 0.

estimates of odds ratios in the population under study. The justification for this claim is beyond the scope of this book, but if readers are willing to accept it at face value, we can proceed with an illustration of how logistic regression models can be used to analyze case-control studies.

In §21.2 we mentioned the importance of age and the possible matching of cases and controls on age in case-control studies. Chapter 5 shows, in the context of 2 × 2 tables, that matching is a special case of stratification. In particular, pair matching, i.e., selecting one specific control for each case, corresponds to the use of strata of size two. Section 11.3 introduced the stratified version of a logistic regression model in the special case of two strata. A more general version of this stratified model is specified by the equation

$$\log\left\{\frac{\Pr(Y=1\,|\,\underset{\sim}{x})}{1-\Pr(Y=1\,|\,\underset{\sim}{x})}\right\}=a_j+\sum_{i=1}^{k} b_i x_i,\ \ j=1,\ \ldots, J \tag{21.5}$$

where j indexes the strata and the b_i's are the logarithms of odds ratios associated with the covariates in $\underset{\sim}{X} = \{X_1, X_2, ..., X_k\}$. As was the case in §11.3, the b_i's are assumed to be the same for each stratum.

Table 21.3 presents an analysis of the data reported by Weiss et al. [67] when cases and controls are grouped in strata defined by one-year age intervals. The design of a case-control study usually includes some degree of matching on age; this ensures moderate balance between cases and controls in age-defined strata. Stratification on one-year age intervals is not always possible, but most studies would likely support strata based on five-year age intervals. With these one-year age intervals and women aged 50–74 years, problems would arise if we tried to estimate all 24 of the a_j's. The possibility of problems like this, and a method for avoiding them, was alluded to in §11.3. Thus, table 21.3 is based on a conditional analysis of model (21.5). This method of analysis adjusts for the stratification, but provides estimates of the b_i's without having to estimate the a_j's. We will not explain how a conditional analysis is carried out, but simply assure the reader that it does not alter the interpretation of the estimated b_i's which was presented in chapter 11.

The regression vector, $\underset{\sim}{X}$, for each subject in the case-control study of endometrial cancer includes a binary variable, X_1, which is equal to zero unless the subject had a duration of use of exogenous estrogen of between one and eight years ($X_1 = 1$), a secondary variable, X_2, which is equal to zero unless the subject had a duration of estrogen use in excess of eight years ($X_2 = 1$), and additional indicator variables for obesity, hypertension and parity. On the basis of calculations which were outlined in §11.3, we conclude that estrogen use of between one and eight years is associated with an estimated odds ratio for endometrial cancer of exp(1.37) = 3.94; the associated 95% confidence interval for the odds ratio is (2.46, 6.30). The corresponding estimate of the odds ratio and confidence interval associated with eight years or more of estrogen use are exp(2.60) = 13.46 and (8.25, 21.98), respectively.

21.5. Confounding and Effect Modification

In the epidemiological literature, the terms confounding factor and effect modifying factor receive considerable attention. An epidemiological study is frequently designed to investigate a particular risk factor for a disease, and it is common practice to refer to this risk factor as an exposure variable, i.e., exposure to some additional risk. A confounding factor is commonly considered to be a variable which is related to both disease and exposure. Variables which have this property are discussed in §5.2 and will tend to bias any estimation of the relationship between disease and exposure. An effect modifying factor is a variable which may change the strength of the relationship between disease and exposure. The odds ratio, which associates exposure and disease, would vary with the level of an effect modifying variable. Confounding variables, and to a lesser extent effect modifying variables, are treated somewhat differently than exposure variables of direct interest because the former tend to be factors which are known to be related to disease. Therefore, it is necessary to adjust for confounding and effect modifying factors in any discussion of new risk factors. Although in formal statistical procedures this distinction between variables is not made, it can be an important practical question. Consequently, we propose to indicate how these concepts can be viewed in the context of regression models for case-control studies.

The analysis of epidemiological data using regression models like (21.3) and (21.5) will only identify something we choose to call 'potential' confounding variables. This is because the interrelation of the covariates is irrelevant in such an approach. Thus, the relationship between exposure and a potential confounding variable is never explored. Provided that the regression model is an adequate description of the data, its use will prevent a variable which is confounding in the common epidemiological sense from biasing the estimation of the odds ratio.

Effect-modification in a logistic regression model corresponds to something called an interaction effect in the statistical literature. Sections 15.3 and 15.4 contain an extensive discussion of the notion of interaction in an analysis of variance setting. Additional occurrences of interaction terms in regression models that we fit to data may be found in §§11.5 and 12.3. Although the use of the term effect modifier is a historical fact, and is unlikely to disappear quickly from the epidemiological literature, Breslow and Day [68] indicate that 'the term is not a particularly happy one however'.

It is wise to regard most regression models as convenient, empirical descriptions of particular data sets. An interaction effect is a concept which only has meaning within the context of a particular model. If an interaction is identified, the aim of subsequent analyses, which may involve alternative regression models, should be to understand the nature of the data and of the biological process which has resulted in the empirical description obtained.

For example, table 21.4 adds two additional variables to the model summarized in table 21.3. The variable X_1X_4 is an interaction term which is the product of X_1 and X_4; X_1X_4 is one only if a woman is hypertensive and used estrogen for one to eight years, and is zero otherwise. A second interaction term, X_2X_4, identifies women who are hypertensive and who have used estrogen for more than eight years ($X_2X_4 = 1$).

According to the analysis summarized in table 21.4, the coefficient associated with X_1X_4 is significantly different from zero. Thus, estrogen use of one to eight years would have an estimated odds ratio of $\exp(1.78) = 5.93$ for non-hypertensive women and $\exp(1.78 - 1.04) = 2.10$ for hypertensive women. This indicates that the effect of moderate estrogen exposure is modified by hypertensive status. Alternatively, if we adopt the approach described in §13.3, we can interpret this significant interaction as indicating that exp(1.78 + 0.60 –

Table 21.4. The results of an odds ratio regression analysis of endometrial cancer incidence incorporating interaction terms for hypertension and estrogen use (cf. table 21.3); the analysis stratifies on baseline age

Risk factor	Regression variable	Definition	$\hat{b}$	SE[a]	Significance level[b]
Estrogen use	X_1	1 if duration of use between 1 and 8 years; 0 otherwise	1.78	0.37	<0.0001
	X_2	1 if duration of use 8 years or greater; 0 otherwise	2.40	0.35	<0.0001
Obesity	X_3	1 if weight greater than 160 lbs; 0 otherwise	0.75	0.33	0.02
Hypertension	X_4	1 if history of high blood pressure; 0 otherwise	0.60	0.31	0.05
Parity	X_5	1 if number of children 2 or greater; 0 otherwise	0.58	0.31	0.06
Interaction terms	X_1X_4	1 if both X_1 and X_4 equal 1; 0 otherwise	–1.04	0.50	0.04
	X_2X_4	1 if both X_2 and X_4 equal 1; 0 otherwise	0.53	0.53	0.32

Maximized log-likelihood = –283.08.

[a] Estimated standard error of $\hat{b}$.

[b] Estimated significance level for testing the hypothesis b = 0.

1.04), the individual odds ratio associated with being hypertensive and a moderate duration estrogen user, is less than the product of the odds ratios associated with hypertension, i.e., exp(0.60), and estrogen use of one to eight years, i.e., exp(1.78), separately. In any event, it appears that estrogen use of one to eight years duration does not further increase a hypertensive woman's risk of endometrial cancer.

The arbitrary grouping of the duration of estrogen use which appears in tables 21.3 and 21.4 may not provide the simplest empirical model. Table 21.5 presents a model where estrogen use is defined to be the logarithm of (estrogen duration use + 1); the constant value 1 is added to prevent infinite values of the covariate. This model fits the data of Weiss et al. [67] as well as the model summarized in table 21.4. That this is the case can be determined by comparing maximized log-likelihoods, a technique which was mentioned briefly at the end of §16.6. The formal details of the assessment are a bit more complicated

Table 21.5. The results of an odds ratio regression analysis of endometrial cancer incidence which models estrogen use with a continuous covariate (cf. table 21.4); the analysis stratifies on baseline age

Risk factor	Regression variable	Definition	$\hat{b}$	SE[a]	Significance level[b]
Estrogen use	X_1	logarithm of (duration of use + 1)	0.98	0.13	<0.0001
Obesity	X_3	1 if weight greater than 160 lbs; 0 otherwise	0.67	0.33	0.04
Hypertension	X_4	1 if history of high blood pressure; 0 otherwise	0.45	0.31	0.15
Parity	X_5	1 if number of children 2 or greater; 0 otherwise	0.59	0.31	0.06
Interaction	X_1X_4	X_1 if X_4 equals 1; 0 otherwise	–0.02	0.18	0.91

Maximized log-likelihood = –286.543.

[a] Estimated standard error of $\hat{b}$.

[b] Estimated significance level for testing the hypothesis b = 0.

here. However, it can be seen that the analysis presented in table 21.5 requires only one variable to model the effect of estrogen use, and the interaction between estrogen use and hypertension is no longer significant. Thus, an interaction which was present in one empirical description of the data need not appear in another.

21.6. Mantel-Haenszel Methodology

Historically, and for convenience, many epidemiological studies have concentrated on a binary exposure variable and examined its relationship to disease after adjustment for potential confounding variables. Therefore, the analysis of such studies has often been based on stratified 2 × 2 tables of the type discussed in chapter 5 and shown in table 21.6.

In chapter 11 and in this chapter, we have indicated how logistic regression models can be used to estimate odds ratios from data of this type. However, the use of logistic regression models does require a reasonably sophisticated computer package, and simpler methods of estimation have a long history of

Table 21.6. A 2 × 2 table summarizing the binary data for level i of a confounding factor

	Confounding factor level i		Total
	success	failure	
Group 1	a_i	$A_i–a_i$	A_i
Group 2	b_i	$B_i–b_i$	B_i
Total	r_i	$N_i–r_i$	N_i

use. The most widely used estimate of ψ, the common odds ratio, based on stratified 2 × 2 tables is the Mantel-Haenszel estimate, for which the defining equation is

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k} a_i(B_i - b_i)/N_i}{\sum_{i=1}^{k} (A_i - a_i)b_i/N_i}, \qquad (21.6)$$

where there are k distinct 2 × 2 tables of the type illustrated in table 21.6. Although it may not be apparent from equation (21.6), this estimate is a weighted average of the individual odds ratios $\{a_i(B_i – b_i)\}/\{(A_i – a_i)b_i\}$ in the k 2 × 2 tables. For completeness, the weight for the ith table is $(A_i – a_i)b_i/N_i$, a quantity which approximates the inverse of the variance of the individual estimate of the odds ratio in the ith table when ψ is near 1. Note also that $\widehat{OR}_{MH}$ is easy to compute and is not affected by zeros in the tables. Research has shown that the statistical properties of this estimate compare very favourably with the corresponding properties of estimates which are based on logistic regression models.

In chapter 5, we introduced the summary χ^2 statistic for testing the independence of exposure and disease. In the epidemiological literature, this same statistic is often called the Mantel-Haenszel χ^2 statistic for testing the null hypothesis that the common odds ratio is equal to 1. There is an extensive statistical literature concerning variance estimates for $\widehat{OR}_{MH}$. Perhaps the most useful estimate is one proposed by Hauck [69] for use in situations when the number of tables would not increase if the study was larger. Its formula, which is not all that simple, is

$$V = \widehat{OR}_{MH} \sum_{i=1}^{k} S_i^2/W_i$$

$$\text{where } S_i = \frac{(A_i - a_i)b_i}{N_i} \text{ and } W_i = \left[\frac{1}{a_i + 1/2} + \frac{1}{A_i - a_i + 1/2} + \frac{1}{b_i + 1/2} + \frac{1}{B_i - b_i + 1/2}\right]^{-1}.$$

Table 21.7. The Doll-Hill [71] cohort study of smoking and coronary deaths in British male doctors

Age range	Smokers		Non-smokers	
	deaths	person-years	deaths	person-years
35–44	32	52,407	2	18,790
45–54	104	43,248	12	10,673
55–64	206	28,612	28	5,710
65–74	186	12,663	28	2,585
75–84	102	5,317	31	1,462

The fraction ½ is added to each of the denominators in W_i to avoid division by zero. Robins et al. [70] discuss other variance estimates that can be used in a wide variety of situations and also are not too difficult to compute.

Because of its good statistical properties, the Mantel-Haenszel estimator of a common odds ratio, ψ, can be recommended for use by researchers with few resources for the complicated calculations of logistic regression. However, readers should remember that an analysis which is based on a logistic regression model is equivalent to a Mantel-Haenszel analysis, and offers the additional generality and advantages of regression modelling.

21.7. Poisson Regression Models

The Doll-Hill [71] cohort study of smoking and coronary deaths among British male doctors was a landmark epidemiological investigation. Table 21.7 provides data from this study as recorded by Breslow and Day [23] and also used by McNeil [72]. The data consist of the number of deaths tabulated by smoking status and age group. In addition, for each smoking status and age group category, the number of years that smoking and non-smoking doctors in the study were observed while belonging to the various age groups are listed. This information is called person-years of follow-up. For example, a smoking doctor who entered the study at age 42 and died or was lost to follow-up at age 57 would belong to the age group 35–44 years for the first two years of follow-up, then to the age group 45–54 for ten years, and finally to the age group 55–64 for three years. This doctor would contribute 2, 10, 3, 0 and 0 person-years to the five successive entries tabulated in column three of table 21.7. Observed deaths in the study are recorded in columns two and four of the table in the row corresponding to the doctor's age at death.

The outcome or response variable for this study, which we will denote by Y, is the number of deaths due to coronary heart disease. The explanatory variables of interest are smoking status, which is used to define a binary variable (0 identifies non-smokers, and 1 denotes smokers), and the ten-year age groups. Since this age group information is a categorical variable with five levels, four binary variables identify the four oldest age ranges and compare each group with the youngest age range (35–44), which would correspond to all four binary variables being coded 0. Readers may recall that we previously encountered the use of k – 1 indicator variables to encode a categorical explanatory variable with k levels in §16.4 and also in chapter 15.

Since the response variable is a count, Poisson regression is a natural method to use in analyzing these data. In chapter 12, we identified the model equation for Poisson regression as

$$\log\lambda = a + \Sigma_{i=1}^{k} b_i X_i ,$$

where λ is the expected or mean count from a Poisson distribution. This regression equation allows the expected counts to depend on the explanatory variables. However, in table 21.7 the counts in columns 2 and 4 depend not only on the explanatory variables but also on the number of person-years of observation that correspond to the observed counts. Therefore, we need to modify the model equation for Poisson regression so that the expected counts also reflect this dependence on person-years.

The necessary modification is equivalent to modelling the event rate, which corresponds in the Doll-Hill study to the death rate rather than the expected counts. If PY denotes the number of person-years corresponding to a particular death count, Y, then the death rate, i.e., number of deaths per person-year of follow-up, is Y/PY. The resulting modified model equation is

$$\log(\lambda/PY) = a + \Sigma_{i=1}^{k} b_i X_i .$$

With this modification, using Poisson regression will be sensible and the results of any analyses will have the same form as we previously described in chapter 12.

Table 21.8 presents the results of fitting two Poisson regression models of this modified type to the data summarized in table 21.7. The model labelled **a** includes only smoking status as an explanatory variable, and demonstrates the strong relationship between smoking status and the physician death rates due to coronary heart disease. However, in this study, as in many cohort studies, age is an important determinant of mortality. Therefore, the analysis labelled **b** summarizes the results of a Poisson regression in which both smoking status and physician age are represented in the model equation by the five binary explanatory variables X_1 and X_2 through X_5, respectively. The estimated regres-

Table 21.8. The results of two Poisson regression analyses of the relationship between coronary deaths and smoking status in British male doctors

Model	Explanatory variable	Estimated regression coefficient	Estimated standard error	Significance level
a	Without adjusting for age			
	a	−5.96	0.10	
	Smoker	0.54	0.11	<0.001
b	Adjusting for age			
	a	−7.92	0.19	
	Smoker	0.36	0.11	0.0005
	Age 45–54	1.48	0.20	<0.001[1]
	55–64	2.63	0.18	
	65–74	3.35	0.19	
	75–84	3.70	0.19	

[1] Likelihood ratio test.

sion coefficients associated with membership in one of the ten-year age groups in the study are significantly different from zero, highlighting the important role that age plays in coronary deaths. Although the regression coefficient associated with smoking status in the latter analysis, i.e., **b**, is slightly smaller than the corresponding estimate obtained without adjusting for age, there is still strong evidence against the hypothesis of no relationship between smoking status and coronary deaths. By including the age group information in the regression model, we know that the importance of smoking status cannot be attributed to any confounding between smoking and age, since the information about age is included in the model when we test the hypothesis that the regression coefficient associated with smoking status is equal to zero.

In the context of this study, the relative rates of events which are specified by a Poisson regression model can be interpreted as relative risk. The event of interest in the Doll-Hill study is death due to coronary heart disease, and the mean of the assumed Poisson distribution can be interpreted as the expected death rate. Therefore, if b_j is the regression coefficient associated with a particular binary explanatory variable, such as smoking status, $\exp(\mathrm{b}_j)$ represents a ratio of death rates, one for smokers and one for non-smokers, i.e., the relative risk of death associated with smoking. Thus, the key feature of the analysis that we can extract from the age-adjusted results in table 21.8 is the estimated relative risk of coronary death associated with smoking, adjusted for age, which is $\exp(0.36) = 1.43$. And if we use the estimated standard error for $\hat{\mathrm{b}}_1$, which is

Table 21.9. Values of the explanatory variable, X_6, representing an approximate interaction between age and smoking status for the Poisson regression analysis of the Doll-Hill cohort study

Smoking status	Age group, years				
	35–44	45–54	55–64	65–74	75–84
No	1	0.5	0	–0.5	–1
Yes	–1	–0.5	0	0.5	1

0.11, to derive the 95% confidence interval 0.36 ± 1.96(0.11), i.e., (0.14, 0.58), for b_1, then a corresponding 95% confidence interval for the relative risk associated with smoking is (exp(0.14), exp(0.58)) or (1.15, 1.79).

Readers will notice that in table 21.8 we have only cited one p-value in the column labelled 'Significance level' for the set of four explanatory variables that encode the information about age. Although each of the four regression coefficients for age could be examined separately, the hypothesis summarizing the notion that a physician's age is not an important systematic effect in the regression model is best expressed as $b_2 = b_3 = b_4 = b_5 = 0$, simultaneously. To investigate the plausibility of this joint hypothesis, we use a likelihood ratio test, which we previously discussed in §16.6. The observed value of the test statistic for these coronary mortality data is 893.8, which is calibrated against the sampling variability summarized in the χ^2 distribution with four degrees of freedom. As the results for the age-adjusted Poisson regression indicate, the significance level of this test that $b_2 = b_3 = b_4 = b_5 = 0$ is very small, underscoring the importance of adjusting for age in our analysis of the relative risk of coronary mortality associated with being a male doctor who smokes.

As well as being a potential confounding factor, age would be an effect modifier if the data collected by Doll and Hill provided evidence that the relative risk of smoking somehow depended on age.

To investigate this possibility requires the exercise of a little skill, since the age-adjusted regression model, which is summarized in table 21.8, already involves six explanatory variables with corresponding estimated regression coefficients. In total, we only have 10 independent observations; they are the observed numbers of deaths in the various subject groups defined by smoking status and the five ten-year age groups. However, with a moderate degree of creativity, we can introduce a simple smoking-age interaction via a single extra explanatory variable, X_6. The possible values for this new variable are given in table 21.9, and depend on age and smoking status.

Table 21.10. The results of a Poisson regression model that includes an interaction between smoking status and age

Model	Explanatory variable	Estimated regression coefficient	Estimated standard error	Significance level
c	Including age as an effect modifier			
	a	–8.28	0.24	
	Smoker	0.52	0.13	<0.001
	Age 45–54	1.58	0.20	<0.001[1]
	55–64	2.84	0.20	
	65–74	3.68	0.22	
	75–84	4.11	0.24	
	X_6	–0.31	0.10	0.001

[1] Likelihood ratio test .

Assigning these values amounts to adopting a linear measurement scale with the values 1, 2, 3, 4 and 5 for the five increasing ten-year age groups, and an indicator variable for smoking status which equals 0 if a doctor was not a smoker, and 1 otherwise. The numerical value of X_6 represents the interaction of smoking status and age group, and is calculated by first centering each component – age group around the value 3 and smoking status around 0.5 – and then calculating their product to get the values shown in the table. By centering each component in the product (interaction) variable, we reduce the correlation between the explanatory variable X_6 and the corresponding main effect variables X_1 through X_5, and should thereby improve the numerical properties of the fitted model.

Table 21.10 summarizes the results of fitting a third Poisson regression model, labelled **c**, which adds the interaction term X_6 to the age-adjusted regression model, i.e., **b**, in table 21.8. The coefficient for this interaction term represents a measure of the effect-modifying aspects of age on the risk of coronary death associated with smoking.

The results displayed in table 21.8 indicate that both smoking status and age are important explanatory variables, not only statistically but also in terms of the size of the epidemiological effects. Moreover, as the estimated regression coefficient for X_6 and its corresponding standard error indicate (see table 21.10), age is not just a potential confounding factor with respect to coronary mortality; it also appears to be an effect modifier. This conclusion is supported by the significance level of the hypothesis test that b_6 is zero. Since the estimated value of b_6 is noticeably different from zero, the effect of a doctor's

Table 21.11. The estimated relative risks of coronary mortality by smoking status and ten-year age bands, based on the Poisson regression model summarized in table 21.10

Age	Smoking status	
	no	yes
35–44	1.00	3.12
45–54	5.67	13.0
55–64	23.3	39.1
65–74	62.9	77.7
75–84	113.4	102.8

smoking status on mortality due to coronary heart disease depends systematically on the age group to which he belongs. In particular, since the estimated value of b_6 is negative, the relative risk associated with smoking declines with age.

Perhaps the best summary of the results of our Poisson regression analysis involving smoking status, age, and the modelled interaction between these two explanatory variables is a set of estimated relative risks. Table 21.11 shows these estimates for smokers and non-smokers in the different age groups. The values are all compared to the experience of doctors in the 35–44 years age group who didn't smoke.

Although this introduction to a regression analysis of the Doll and Hill cohort study of coronary mortality has involved some modelling subtleties that not every reader may have entirely grasped, we hope everyone has been able to appreciate the potential usefulness of Poisson regression methods in analyzing cohort studies.

21.8. Clinical Epidemiology

At the beginning of this chapter, we remarked that previous chapters were primarily concerned with clinical studies, whereas epidemiology concentrates on the incidence of disease. There is also now increasing use of the term 'clinical epidemiology'.

Certain areas of clinical investigation, such as diagnostic tests which we will discuss in chapter 22, are sometimes particularly associated with clinical epidemiology. This term is also used by some writers to refer to the application of epidemiological methods to clinical medicine. However, we are sympathetic to the more general definition given by Weiss [73]. Weiss argues that 'epidemiology is the study of variation in the occurrence of disease and of the reasons for that variation'. He defines clinical epidemiology in a parallel way as 'the study of variation in the *outcome* of illness and of the reasons for that variation'.

Whether a specific term for this activity is required may be debatable, but whatever one's reaction, semantically, to the term clinical epidemiology, the area of study defined by Weiss is of obvious importance and would encompass many of the examples used in other chapters of this book. We trust, therefore, that this book will be useful to readers who regard themselves as interested in clinical epidemiology.

22

Diagnostic Tests

22.1. Introduction

In view of the widespread use of diagnostic tests in the practice of modern medicine, it is probably safe to assume that most readers have some acquaintance with their use. However, the underlying concepts pertaining to diagnostic tests, and to their use for diagnosis or screening purposes, are often less familiar, and perhaps not well understood. Because various tests, and the results derived when they are used, frequently play a vital role in the total diagnostic process, we believe that a modest effort spent in understanding diagnostic testing, from a statistical perspective, will pay worthwhile short- and long-term dividends. To ensure that each reader has the same, sound basis on which to build his or her grasp of this key element of modern practice, we will try to assume nothing and begin our exposition with some rudimentary ideas.

22.2. Some General Considerations

Figure 22.1 displays a plot of paired enzyme-linked immunosorbent assay (ELISA) measurements derived from a single aliquot of blood collected from each of 1,762 blood donors. In the latter half of the 1980s and throughout the following decade, ELISA tests were widely used to screen blood donations for the presence of antibodies to the Type 1 human immunodeficiency virus (HIV-1). Although a report received from the testing laboratory might simply indicate that the specimen tested was either ‘Reactive’ or ‘Non-reactive’, this binary outcome was a classification, based on protocols provided by the test manufacturer, of the actual ELISA test measurement evaluated in the lab. The

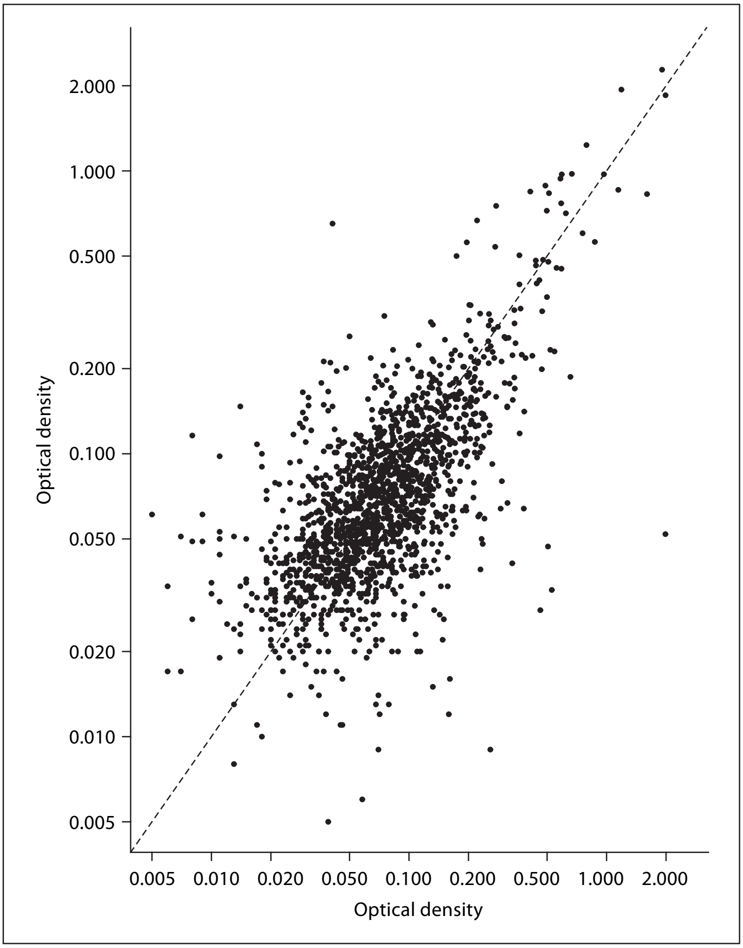


Fig. 22.1. A scatterplot of paired optical density measurements from repeated ELISAs using specimens from 1,762 blood donors.

measurements displayed as points on the plot represent two separate evaluations of the same physical characteristic of each specimen, which happens to be the optical density of whatever was left after four possible chemical reactions, in an ordered sequence of 15 steps, had been initiated via the testing procedure.

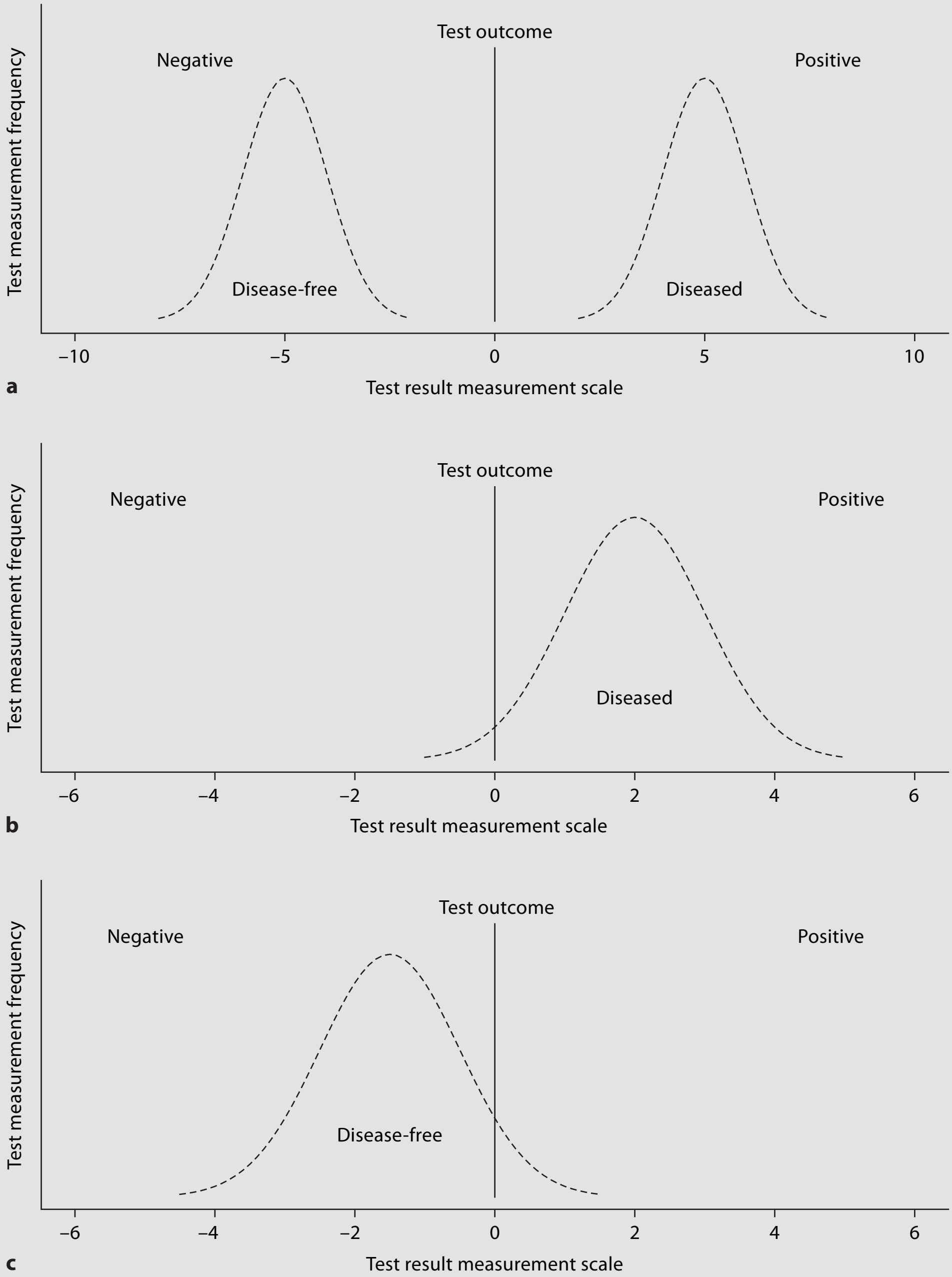
Because the optical density measurements obtained during this process are at least partially determined by factors specific to the microplate on which the specimen is assayed, it would be unrealistic to expect that two optical density measurements derived from the same specimen, but assayed on different ELISA plates, would necessarily be identical. However, if the two measurements obtained from the same specimen were assayed on the same ELISA plate, it seems reasonable to suppose that the resulting optical densities would be quite similar. In that case, a plot like the one displayed in figure 22.1 should show that virtually all of the points displayed lie very close to the reference line of equal values running from the lower left corner to the upper right corner of the diagram. If repeated measurements of the same characteristic of a fixed specimen or unit, obtained under the same conditions, have virtually the same numerical value, they are said to be repeatable. Since the optical density measurements for each specimen represented in figure 22.1 were derived from two ELISAs evaluated on the same plate, the testing kit clearly lacked repeatability. Incidentally, a logarithmic scale was used on each axis of the plot in order to enhance simultaneous visualization of all 1,762 paired optical density measurements.

Fortunately, test characteristics such as measurement repeatability and reproducibility are investigated during the licensing or approval process of an appropriate regulatory authority such as the US Food and Drug Administration. Clinical chemistry, which is probably not appreciated as much as it deserves, plays a crucial role in first developing, and then improving and maintaining, the complex measurement systems that underly virtually all of the diagnostic tests used in modern clinical practice.

As we previously indicated, many diagnostic test results are not reported as measurements like the optical density values displayed in figure 22.1, but rather as binary outcomes, e.g., reactive/non-reactive, normal/abnormal, positive/negative or perhaps diseased/disease-free. This classification of the underlying physical measurement is based on characteristics of the measurement system that the manufacturer has incorporated into the test protocol. If we suppose that the population in which the test is approved for use can be subdivided into the two groups identified by the binary outcomes, say diseased and disease-free, then individuals in the diseased group will have a distribution for their test measurements.

For convenience, we will assume that this distribution looks like the normal distribution. Likewise, the individuals who are disease-free will have a separate distribution for their test measurements. If it also resembles the normal distribution, we might have a situation similar to the one depicted in the upper panel of figure 22.2. Readers can no doubt see immediately that this

Fig. 22.2. A pictorial framework for diagnostic testing. **a** The ideal world; **b**, **c** more realistic settings.


Test outcome
Negative
Positive
Disease-free
Diseased
Test measurement frequency
−10
−5
0
5
10
a
Test result measurement scale
Test outcome
Negative
Positive
Diseased
Test measurement frequency
−6
−4
−2
0
2
4
6
b
Test result measurement scale
Test outcome
Negative
Positive
Disease-free
Test measurement frequency
−6
−4
−2
0
2
4
6
c
Test result measurement scale

would be an ideal world, because test measurements that turned out to be positive would always correspond to diseased individuals (so-called true positives), and test measurements that were negative values would indicate persons that were disease-free (so-called true negatives). Fortuitously, the sign of the test measurement would suffice to classify subjects without ever making an error.

Regrettably, the world of diagnostic testing is rarely so unequivocally ordered. For example, consider the prostate-specific antigen (PSA) test that is routinely used to screen for prostate cancer. Most men who are middle-aged or older will have some detectable antigen in their blood, say 0.5 ng/ml, and those individuals who have advanced prostate cancer will have high concentrations, perhaps in excess of 20 ng/ml. However, as Catalona et al. [74] report, a concentration of 6.8 ng/ml may be observed in an individual who is disease-free, or in someone with early cancer. As we have depicted in figure 22.2b and c, the distributions of test measurements in the diseased and disease-free groups overlap to some extent. Regardless of where the test outcome threshold is situated on the measurement scale, some diseased individuals will be incorrectly classified as 'Negative', and hence give rise to the kind of diagnostic error known as a false-negative outcome. Likewise, disease-free individuals whose test measurement is larger than the outcome threshold (see fig. 22.2c) will be identifed as 'Positive', which represents the other kind of diagnostic error – one known as a false-positive outcome. And provided the distributions of test measurements derived from individuals in the diseased and disease-free groups overlap, neither type of diagnostic error can be avoided, or eliminated. Because the diagnostic test has a single outcome threshold, moving it to the right will reduce the false-positive error rate, but simultaneously increase the false-negative error rate. Similarly, moving the single outcome threshold to the left will reduce the false-negative error rate; however, the false-positive error rate automatically increases. Only by changing the distributions of test measurements in the diseased and disease-free groups, so that they begin to approximate what is depicted in figure 22.2a, can investigators make simultaneous reductions in the rates of both types of diagnostic test errors.

22.3. Sensitivity, Specificity, and Post-Test Probabilities

Many readers will recognize that the true status of a patient, and the outcome of a diagnostic test for the presence of a particular condition or disease, can be represented conveniently in the cells of a 2 × 2 table, such as the one depicted in table 22.1. The probability that a diseased individual is correctly identified as 'Positive' is often called the sensitivity of the diagnostic test. Ob-

Table 22.1. A 2 × 2 table summarizing the four possible outcomes of a diagnostic test for the presence of a condition or disease

Test outcome	True disease status	
	diseased	disease-free
Positive	True positive	False positive
Negative	False negative	True negative

viously, one desirable characteristic of a good diagnostic test is that it has a sensitivity value close to 1. Likewise, the probability that a disease-free individual is correctly identified by the diagnostic test as 'Negative' is often called the test specificity. Another preferred characteristic of a diagnostic test is that it should be highly specific, i.e., have a specificity that is close to 1.

To make these two notions more concrete, we can refer to figure 22.2, especially the middle and lower panels. Since sensitivity represents the probability that a diseased individual is correctly identified as 'Positive', it corresponds to the fraction of the area under the probability curve for test measurements obtained from diseased individuals that lies to the right of the test outcome threshold (see fig. 22.2b). Thus, a highly sensitive test is one for which virtually all of the test measurements in the population of diseased individuals lie on the side of the test outcome threshold designated as a positive (or diseased) outcome. The remaining fraction of the area under the same probability curve represents the false-negative error rate. And specificity, which is the probability that a disease-free individual is correctly identified as 'Negative', corresponds to the fraction of the area under the probability curve for test measurements from disease-free individuals that lies to the left of the test outcome threshold (see fig. 22.2c). If virtually all the test measurements that might be obtained in the disease-free population lie on the side of the test outcome threshold corresponding to a negative (or disease-free) outcome, then the test is said to be highly specific. Of course, the false-positive error rate is equal to the remaining fraction of the area under the same probability curve for measurements from disease-free individuals.

These two characteristics of all diagnostic tests – sensitivity and specificity – are estimated during the approval process. However, it should be noted that the reported estimates may be overly optimistic due to inadequacies in the design of the study through which the estimates were obtained. Ransohoff and Feinstein [75] refer to the 'spectrum' or range of characteristics represented in subjects participating in any study used to investigate the characteristics of a putative diagnostic test. They cite various examples from the medical literature to convince researchers that inadequacies in the pathological, clinical or co-morbid components of the spectrum of diseased and disease-free subjects used to validate a newly-developed diagnostic tool should prompt clinicians to be cautious about apparently promising tools that have not been adequately scrutinized. One such example is a carcinoembryonic antigen (CEA) test for colon cancer. Initial studies reported high sensitivity and specificity – each in excess of 0.90 – but this was apparently due to their having been estimated in patients with extensive disease. For patients with localized disease, the sensitivity of the CEA test was eventually shown/estimated to be as low as 0.37.

Unfortunately, the problems identified by Ransohoff and Feinstein nearly 30 years ago continue to be repeated, prompting Reid et al. [76] to conclude in 1995 that 'most diagnostic tests are still inadequately appraised'. In a review of 112 reports published in the general medical literature between 1978 and 1993, fewer than one study in three was deemed to have provided even a rudimentary description, e.g., age and sex distribution, range of clinical symptoms and/or disease stage, of the patient spectrum used to investigate the potential diagnostic tool.

However, when a physician has a patient's test outcome report in her hand and needs to reach a decision about a particular diagnosis, even knowing that the test she ordered is both highly sensitive and highly specific in the relevant subgroup of patients does not directly address the immediate problem. That is because, unless the sensitivity and specificity are simultaneously 1 – and hence figure 22.2a is apropos – regardless of what is written on the lab report, the outcome could be erroneous.

Instead of referring to the sensitivity and specificity, what our physician is really interested in knowing is the extent to which a positive (or negative) test outcome accurately predicts the true status of her patient, i.e., diseased (or disease-free). This value or rate is commonly referred to as the post-test probability of disease, or the predictive value of a positive test outcome; if the test outcome is negative, then it is the post-test probability of being disease-free, or the predictive value of a negative test outcome. And these values depend on not only the sensitivity and specificity, but also on a third probability, which is known as the pretest probability or prevalence of the disease or condition.

Although there is a mathematical result, known as Bayes' theorem, that connects sensitivity, specificity and prevalence to the post-test probability of a positive test outcome, we believe it is easier to grasp the sense of this relationship directly. Consider the following example. If a female patient has recently indicated a desire to become pregnant, then a positive pregnancy test result is fairly likely to be a true-positive result. In effect, the pretest probability of a positive pregnancy test is high because of the patient's prior indication. Thus, when the lab report from her pregnancy test comes back positive, both the patient and her

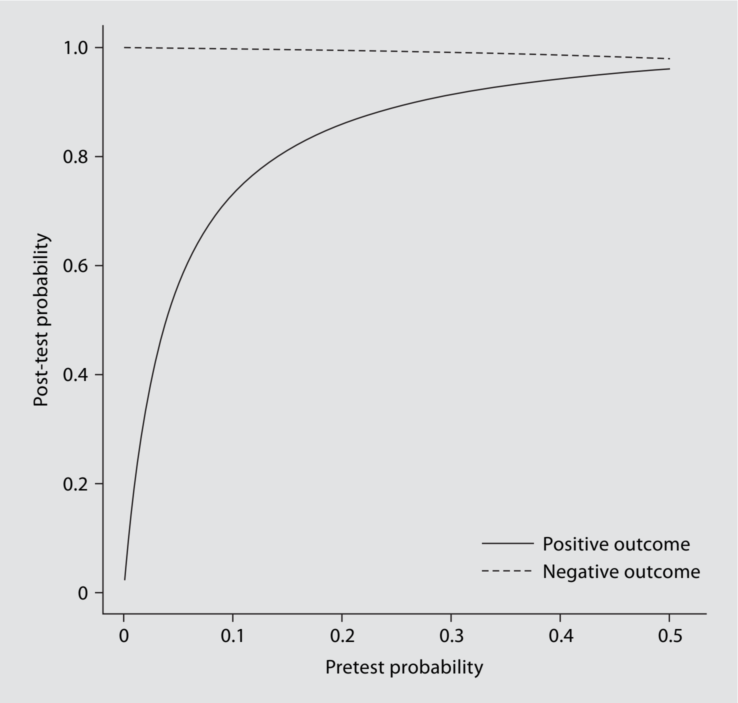


Fig. 22.3. The relationship between pre- and post-test probability that a particular test result is correct; the plots assume a sensitivity of 0.98 and a specificity of 0.96.

physician have little reason to doubt the test result, i.e., the physician believes the test result is a true positive. On the other hand, if the patient and her doctor have recently discussed the choice and use of effective contraceptive practices because she has indicated an aversion to becoming pregnant at the present time, then a positive lab report from a pregnancy test will raise questions in the doctor's mind about the reliability of the test, and may prompt her to request a confirmatory pregnancy test. In this case, the pretest probability of pregnancy is low, because of the recent discussion between the patient and her doctor, and hence the physician has good reason to question the reliability of the test result, i.e., the physician suspects that the test result is a false positive.

In effect, when the pretest probability or prevalence of the disease or condition is substantial, a positive test outcome is probably a correct result, and the post-test probability that a positive test outcome correctly indicates the patient is diseased will be high. The complementary post-test probability that a patient whose test outcome was positive represents a false-positive test out-

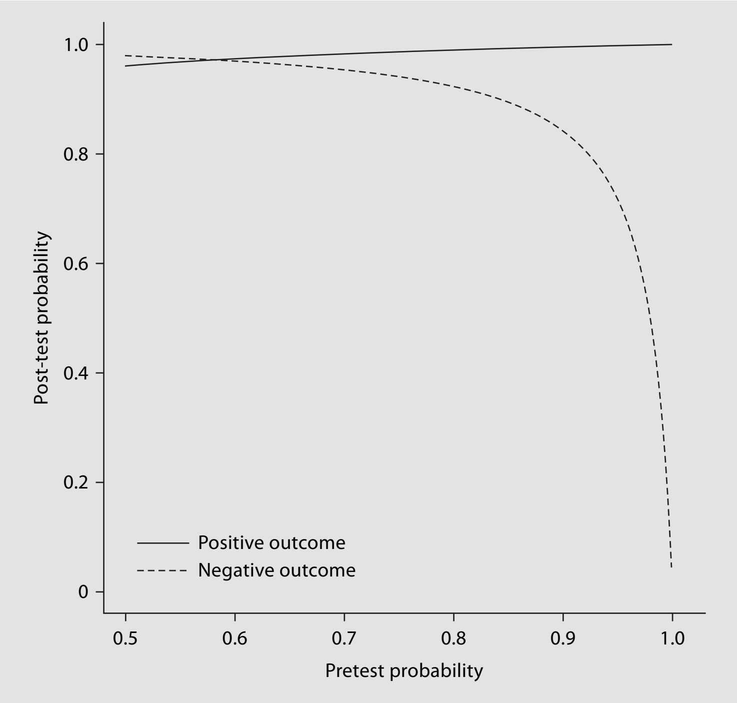


Fig. 22.4. The relationship between pre- and post-test probability that a particular test result is correct; the plots assume a sensitivity of 0.98 and a specificity of 0.96.

come from a disease-free individual will necessarily be low, since the only other explanation for a positive test outcome has a high post-test probability, and these two post-test probabilities associated with a positive test outcome must add to one.

Likewise, when the pretest probability or prevalence is low, a negative test outcome is most likely a correct result, and the corresponding post-test probability that a negative test outcome correctly indicates the patient is disease-free will also be high. However, the complementary post-test probability that a negative test outcome constitutes a false-negative test result from a patient who is diseased will be low, since the only other explanation for a negative test outcome has a high post-test probability, and these two post-test probabilities associated with a negative test result necessarily add to one.

This intuitive perspective is illustrated in figures 22.3 and 22.4, for a diagnostic test with sensitivity and specificity in the relevant patient population that are estimated to be 0.98 and 0.96, respectively. The graph in figure 22.3

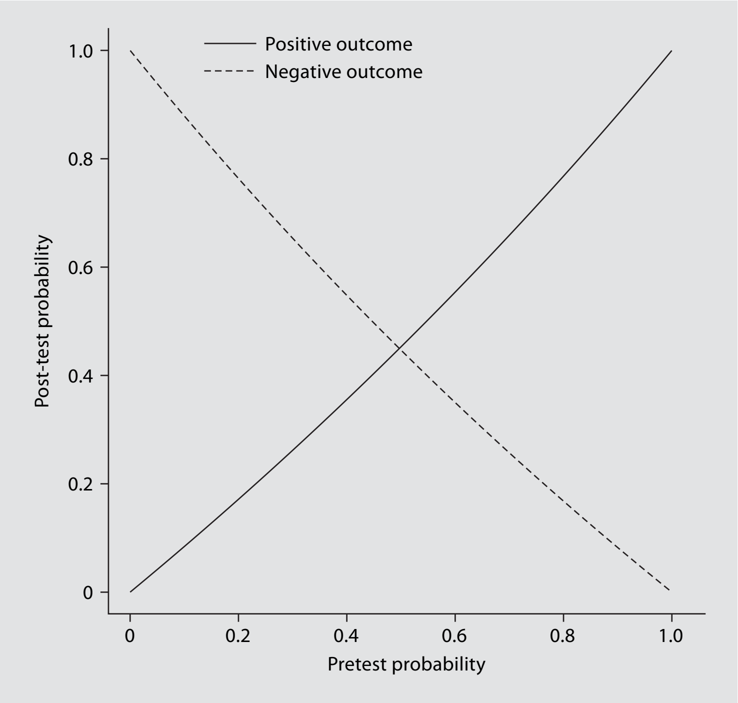


Fig. 22.5. The relationship between pre- and post-test probability that a particular test result is correct; the plots assume a sensitivity of 0.48 and a specificity of 0.42.

shows how the post-test probability that a positive test outcome correctly indicates the presence of disease increases from a very low value, when the pretest probability is virtually negligible, to approximately 0.96 when the pretest probability is 0.5. In the meantime, the post-test probability that a negative test outcome correctly indicates disease-free status is almost a constant value, and very high, i.e., 0.98. However, as the pretest probability of disease increases from 0.5 to 1.0 – see figure 22.4 – the post-test probability that a negative test outcome correctly indicates disease-free status decreases dramatically from 0.98 to approximately 0.05, while the corresponding post-test probability that a positive test result correctly identifies that the patient is diseased is virtually unchanged, and always in excess of 0.96.

On the other hand, if the test sensitivity and specificity are low, so that their combined sum is less than one, the value of a diagnostic test, as encapsulated in the relationship between the pre- and post-test probabilities, is less evident. Figure 22.5 illustrates this for the case of a diagnostic test with an estimated sensitivity of 0.48 and a specificity of 0.42 in the relevant patient population. In this case, one can show mathematically (although we won't attempt to do so here) that because the test characteristics are sufficiently unsatisfactory, knowing the test outcome – either positive or negative – has actually muddied the diagnostic waters. In each case, the post-test probability that the test outcome correctly identifies the patient's status is lower than the corresponding pretest probability.

22.4. Likelihood Ratios and Related Issues

Although the medical literature concerning the use and interpretation of diagnostic tests often refers to sensitivity and specificity, in recent years the term likelihood ratio of a positive test result has become more common. It appears that this terminology was first introduced by Lusted [77], and was subsequently popularized in the 1990s by Sackett et al. [78]. This use of the term likelihood ratio involves a different purpose than that for which it has been used elsewhere in this book, i.e., for the testing of hypotheses. What Lusted called a likelihood ratio corresponds to the relative probability of a positive diagnostic test in a diseased individual compared with a non-diseased individual. Because this terminology is now in common use, it seems advisable to explain more fully what the likelihood ratio of a positive test result represents, and what role it plays in using and understanding diagnostic tests.

In most practical clinical settings, physicians would prefer to order a particular diagnostic test only if the result enables them to rule in or rule out a certain disease. Ruling in the disease would follow if the probability of a true-positive outcome in a diseased individual is considerably more likely than a false-positive error in a disease-free patient. Of course, these are the only two ways in which a positive test outcome could arise. The former probability is the sensitivity of the test, and the latter is the probability of a false-positive outcome, or 1 minus the specificity of the test. It is the ratio of these two probabilities that corresponds to the likelihood ratio of a positive test result. If the goal of the test is to rule in disease, this likelihood ratio should be at least one, and preferably much larger than one. Pictorially, it represents the ratio of the two areas shown in figure 22.2b and c that lie on the positive (right) side of the test outcome threshold.

Likewise, a test that enables a physician to rule out a particular disease in his patient would be one such that the probability of a true-negative outcome in a disease-free individual is substantially larger than the probability of a false-negative error in someone who is diseased. In effect, the specificity of the test is greater, and ideally much greater, than the probability of a false-negative

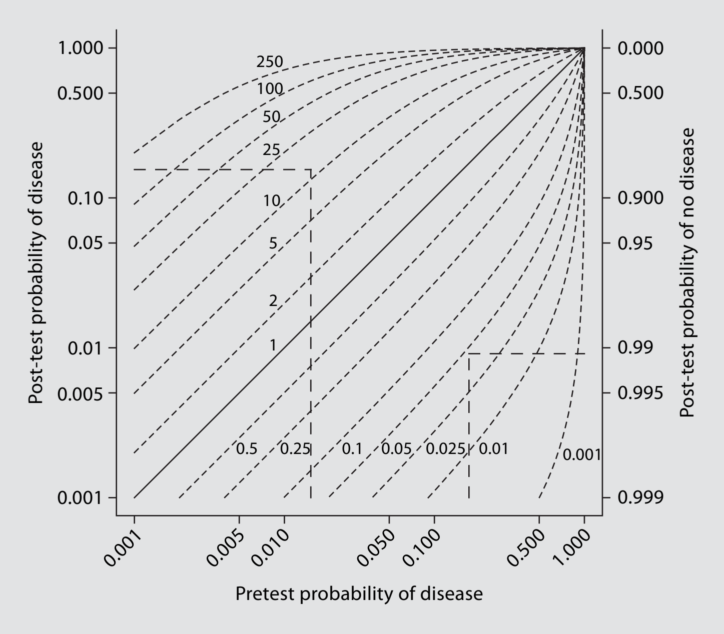


Fig. 22.6. The relationship between pre- and post-test probability that a positive (negative) test result correctly predicts disease (no disease), as a function of the likelihood ratio of a positive (negative) test outcome.

result. The ratio of these two probabilities that are associated with a negative test outcome is often called the likelihood ratio of a negative test result. It corresponds to the ratio of the two areas lying on the negative (left) side of the test outcome threshold shown in figure 22.2b and c. Ideally, this value should also be substantially bigger than one. For reasons of pedagogy, or perhaps consistency of usage, the accepted definition of the likelihood ratio of a negative test result in the medical literature is the reciprocal of the ratio described above, i.e., this likelihood ratio is the probability of a false-negative error divided by the specificity. Consequently, the values of likelihood ratios for negative test outcomes thus defined would typically be less than one, and ideally much smaller than one.

It so happens that if we know the prevalence, or pretest probability, of disease as well as these two so-called likelihood ratios, we can easily calculate the corresponding post-test probability that a positive (negative) test result is correct. Rather than introduce the two specific formulae, we have chosen to present the relationship visually, through the graphs displayed in figure 22.6. The solid diagonal line and the curves plotted with short dashes indicate the explicit conversion of pretest probability to the corresponding post-test probability that a patient whose test outcome is positive is diseased, i.e., their test result correctly indicates their status. The curves plotted with short dashes lying above and to the left of the solid line represent seven different values of the likelihood ratio of a positive test result between 2 and 250, inclusive. Seven additional curves lying below and to the right of the solid line correspond to seven different values of the likelihood ratio of a positive test result between 0.001 and 0.5. The vertical and horizontal dashed lines at a pretest probability of 0.015 and a likelihood ratio of 12 illustrate how to connect an approximate post-test probability of roughly 0.20 with that particular combination of prevalence, i.e., 0.015, and diagnostic test characteristics, i.e., a positive test result likelihood ratio of 12. For example, a test having a sensitivity of 0.60 and a false-positive error rate of 0.05 would have a positive test outcome likelihood ratio of 0.60/0.05 = 12; so also would a test having a sensitivity of 0.96 and a false-positive error rate of 0.08.

Obviously, since the horizontal and (left-hand) vertical scales on the graph are identical, any test with a likelihood ratio of one has a post-test probability that is equal to the pretest probability. For such a diagnostic test, false-positive outcomes are as common as true-positive ones, and the solid line on the graph, which represents a likelihood ratio of 1, links equal values of pretest and post-test probabilities. While such a small likelihood ratio might hardly seem sensible, if the public health consequences of a false-negative test result were sufficiently catastrophic, and provided false-positive outcomes could be suitably identified by some sort of repeat test for the disease or condition, using such a test is not as foolish as it might first seem. In fact, the Guthrie test for congenital hypothyroidism that was used until quite recently to screen all newborn babies in most of the developed world had both a high sensitivity – >0.99 – and a large false-positive error rate, i.e., a small likelihood ratio. Clearly, the public health authorities believed that the stress, for parents, of a repeat blood test to rule out the false-positive results from an initial Guthrie test was vastly outweighed by the benefit of identifying virtually all newborn infants with this treatable condition who would otherwise develop a severe mental handicap.

It is very uncommon for a test to have a likelihood ratio for a positive test outcome that is less than one. However, in this case the likelihood ratio curves plotted in figure 22.6 below and to the right of the solid line allow evaluation of the corresponding post-test probability, which will, in fact, always be smaller than the corresponding pretest value. However, these particular likelihood ratio curves also correspond to preferred values of the likelihood ratio of a negative test outcome and can therefore be used to evaluate approximate post-test probabilities of no disease in a patient whose test outcome is negative.

These values are generated from the pretest probability of disease, which is located on the horizontal axis of the graph, and the likelihood ratio of a negative test outcome; the post-test probability of no disease should be read off the right-hand vertical scale, which is the reverse of its left-hand counterpart. As we remarked above, diagnostic tests that have been selected to rule out disease would typically have likelihood ratios for negative test results that are substantially less than one. By using the dashed curves found below and to the right of the solid line, readers can obtain the approximate value of the probability of no disease in a patient whose test outcome is negative. For example, a test with a specificity of 0.8 and a false-negative error rate of 0.036 has a likelihood ratio for a negative test outcome of 0.036/0.8 = 0.045. If the pretest probability of disease is 0.17, then the pair of vertical and horizontal dashed lines in figure 22.6 located at 0.17 and 0.991, respectively, identify that the post-test probability of correctly ruling out disease in a patient whose test outcome is negative is 0.991, roughly 20% greater than the corresponding pretest value of 1 – 0.17 = 0.83.

Fagan [79] published a nomogram, which is a two-dimensional graphical device, for calculating post-test probabilities from known values for the pretest probability and the relevant likelihood ratio for the diagnostic test. An adapted version of Fagan's nomogram appears in Sackett et al. [78, p. 124], and numerous versions of both devices, both static and dynamic, can easily be located on the internet.

One of the benefits arising from an acquaintance with the notion of diagnostic test likelihood ratios is that we can explore the potential advantage of using the original test measurement that gave rise to a binary reported outcome such as reactive/non-reactive. For example, the serum ferritin concentration, which is a diagnostic test for iron deficiency anaemia, has been extensively investigated as part of a systematic review of tests to diagnose that condition. From detailed information on 2,669 patients, 809 (30%) of whom were iron-deficient, Guyatt et al. [80] estimated the likelihood ratios summarized in table 22.2.

If we assume the prevalence, or appropriate pretest probability, of iron deficiency anaemia is 0.30, we can immediately calculate the post-test probability of disease for each range of serum ferritin concentration test results listed in the table, using either the graph displayed in figure 22.6 or the actual formula that links pretest probability, the six range-specific likelihood ratios, and the corresponding post-test probabilities. The end result of these fairly simple calculations is the final row of entries in table 22.2, i.e., six post-test probabilities associated with the six different intervals that span the entire range of serum ferritin concentration measurements. And on the basis of a serum ferritin concentration measurement for a particular patient, a physician might be bet-

Table 22.2. The relationship between serum ferritin concentration, the likelihood ratio of a positive test outcome, and the post-test probability of disease (when the prevalence is 0.30)

	Serum ferritin concentration, μg/l					
	<15	15–25	25–35	35–45	45–100	≥100
Likelihood ratio	51.9	8.8	2.5	1.8	0.54	0.08
Post-test probability	0.96	0.79	0.52	0.44	0.19	0.03

ter able either to rule in, or possibly rule out, iron deficiency anaemia as the appropriate diagnosis consistent with other signs and presenting symptoms observed in that patient. Or if the diagnosis was still equivocal, perhaps additional, more expensive tests might then be used to zero in on a correct diagnosis of the patient's ailment. Although we will not attempt to explain it here, a physician who is armed with the right information could even identify the post-test probability associated with one diagnostic test result as the pretest probability for the next stage in a series of sequential steps towards a conclusive diagnosis.

The previous discussion of post-test probabilities, and their dependence on the relevant likelihood ratio associated with a positive (or negative) test outcome, may have prompted readers to realize that the problem of identifying an optimal threshold to differentiate between positive and negative test outcomes is not a purely statistical question. Rather, the definition of what is optimal depends on how the test result will be used. In the case of the simple blood test for phenylketonuria, galactosaemia, congenital hypothyroidism, cystic fibrosis and several other conditions that neonates world-wide undergo shortly after birth, the test outcome threshold is deliberately situated to ensure that virtually all infants affected by one or more of these conditions are identified. Although this choice necessarily involves a substantial false-positive rate, additional follow-up ensures that only the affected infants receive appropriate support and treatment for their disease which, thankfully, is quite rare; the prevalence rate for any of these conditions is roughly one infant in 800 births.

A similar situation holds with respect to the protocol used to screen voluntary blood donations for various transfusion-transmitted infectious agents, such as HIV-1 and -2, hepatitis B and C, and syphilis. Since each donated unit is typically tested once, and if that test result is non-reactive then the unit is processed and added to the whole blood inventory, maintaining the safety of the blood supply dictates that the false-negative error rate must be negligible.

The resulting test outcome threshold necessarily involves a substantial false-positive rate for the initial screening of blood donations, and collected units that are identified as reactive are routinely discarded, although follow-up tests (usually two) may be carried out to discriminate between true- and false-positive donors if the blood collection agency has a secondary, diagnostic role in its operational mandate.

As physicians become persuaded of the merits of post-test probabilities, and acquire familiarity with the concept and use of the likelihood ratio of a positive or negative test outcome, clinical investigators are beginning to design research studies that enable them to estimate these key characteristics both for familiar and new diagnostic tools. In doing so, they help steer current and future medical practice towards the goal that was first articulated by Dr. George W. Peabody [81] more than 80 years ago.

> 'Good medicine does not consist in the indiscriminate application of laboratory examinations to a patient, but rather in having so clear a comprehension of the probabilities of a case as to know what tests may be of value … it should be the duty of every hospital to see that no house officer receives his diploma unless he has demonstrated ... a knowledge of how to use the results in the study of his patient.'

23

Agreement and Reliability

23.1. Introduction

Accurate, precise measurement is fundamental to any medical study. It is important to know the extent to which the measurements are subject to error, and the degree to which they meaningfully represent patient status. Likewise, whether the results of a measurement or classification procedure concur in successive applications is essential knowledge. Thus, studies that establish the reliability of any measurement and the agreement among observers who determine it are necessary.

Consider the extremely simple example of the measurement of body temperature. If everyone's body temperature was always the same value, say 37°C, regardless of their health status, there would be little value in ever taking a patient's temperature because that measurement would not help in diagnosing a patient's condition. Variation in body temperature, particularly systematic variation linked to health status, is what makes the body temperature measurement useful in the diagnostic process. However, if that same measurement process exhibited such random variation that the range of observed values in temperature was similar in both healthy and ill patients, then body temperature would cease to be a 'reliable' indicator of ill health. Body temperature is useful because the systematic variation with health status is greater than the random variation associated with the measurement in a particular person. The random variation in body temperature measurement in a particular person results from a combination of factors, one of which is the observer – the person taking the measurement. Thus, the extent of this variation reflects, at least partially, the ability of different observers to 'agree', and hence the sense of confidence that is warranted in the measurement taken by any single observer. This same confidence encompasses the belief that had the measurement been taken by another observer, or indeed repeated by the same observer, the observed temperature would have been similar.

The terms ‘observer reliability’ and ‘observer agreement’ are often used interchangeably; in theory, they are different concepts. Reliability coefficients indicate the ability of the corresponding measurements to differentiate among subjects. For continuous measures, these coefficients are typically ratios of variances: in general, the variance attributed to the differences in the measurement of interest among subjects divided by the total variance in the same measurement. For example, if a particular test was administered at two different time points and the data from those successive determinations were available, in an ideal world the results would be the same. However, variation in the method and location of sampling, as well as variation in other aspects of the measurement process, may give rise to different outcomes. In this context, if both occasions when the test is administered provide the same capability to discriminate between patients with respect to the measurement under investigation, we might claim to have empirical evidence of test measurement reliability.

In contrast, agreement refers to conformity. Agreement parameters assess whether the same observation results if a particular measurement is carried out more than once, either by the same observer or by different observers. Typically, observers are said to exhibit a high degree of agreement if a large percentage of the measurements they carried out actually concurred. Poor agreement corresponds to a situation where observers often made substantially different measurements.

In a more heterogeneous population (with wider ranges of observed measurements), the value of a reliability coefficient will tend to be larger. This reflects the fact that, in heterogeneous populations, subjects are easier to distinguish than in homogeneous populations. In general, one might expect that, in a heterogeneous population, reliability and agreement measures will correspond well. On the other hand, in homogeneous populations reliability is generally low since it is difficult to distinguish between patients; however, agreement may well be high.

There is a considerable literature about assessing reliability and agreement, and that literature reflects different opinions on the best way to characterize these two concepts. Here we adopt the limited aim of presenting some methods that are commonly used to study reliability and agreement, and discussing their various properties. In doing so, we hope to highlight some important issues for the reader.

Table 23.1. BILAG scores calculated for eight lupus patients evaluated by eight different physicians

Patient	Physician								Patient mean
	1	2	3	4	5	6	7	8	
1	10	11	19	13	11	13	13	13	12.88
2	15	19	13	17	19	23	27	11	18.00
3	4	4	1	4	1	19	1	4	4.75
4	5	3	4	5	3	5	4	2	3.88
5	7	18	7	7	7	12	9	9	9.50
6	2	7	3	6	6	7	5	6	5.25
7	8	17	16	9	0	8	5	1	8.00
8	7	18	16	10	8	9	10	7	10.63
Physician mean	7.25	12.13	9.88	8.88	6.88	12.00	9.25	6.63	

23.2. Intraclass Correlation Coefficient

Consider first the situation when the measurement of interest, Y, is a continuous variable. For each of N subjects, we assume there are J measurements made by J different observers. Table 23.1 displays an example of such measurements of a total disease activity score, called BILAG, for patients with systemic lupus erythematosus. In this example, eight patients with lupus were each examined by eight physicians who completed activity questionnaires from which the actual BILAG values were calculated. Thus, for this example, N and J happen to have the same value; however, their equality is not a universal requirement in measurement reliability studies.

It is helpful to represent this measurement study involving N subjects and J observers in terms of the statistical model

$$Y_{ij} = a + b_i + c_j + e_{ij},$$

which is simply another way of representing an analysis of variance model like those we previously encountered in chapter 15. Here we need to use subscripts to be very explicit about the model. The subscript i indexes the N subjects and j indexes the J observers. There are NJ observations in total, J for each of the N subjects, and the subscript pair, ij, uniquely identifies a single observation. The symbols $b_1, ..., b_N$, which are called subject effects, are similar to regression coefficients associated with the subjects labelled 1, ..., N. Likewise, the symbols $c_1, ..., c_J$ – so-called observer effects – are akin to regression coefficients that correspond to observers 1, ..., J. The symbol a is the analog of an intercept term,

Table 23.2. ANOVA table for BILAG reliability/agreement experiment

Term	Sum of squares	DF	Mean square	F-ratio	Significance level
Subject	1,265.61	7	180.80	11.74	<0.001
Observer	261.86	7	37.41	2.43	
Residual	754.76	49	15.40		
Total	2,282.23				

and e_{ij} represents the residual value that makes the left- and right-hand sides of the model equation equal to each other.

Table 23.2 is the ANOVA table that corresponds to fitting this model to the measurement study data summarized in table 23.1.

The N patients involved in such a measurement reliability study are assumed to be a random sample of possible subjects on whom the measurements of interest might have been observed. The 'effect' associated with the subject labelled i, i.e., b_i, is assumed to be one observation from a normal distribution with mean zero and variance σ_b^2. Thus, σ_b^2 represents the variation in the measurement of interest by the same observer from subject to subject. In addition, in the context of measurement agreement and reliability, it is common to think of the J observers as a random sample of possible observers, and to assume that the 'effect' associated with observer j, i.e., c_j, is one observation from a normal distribution with mean zero and variance σ_c^2. Then σ_c^2 represents the variation in the measurement of interest from observer to observer when measuring the same subject. Finally, the NJ residuals, which are represented by e_{ij}, are assumed to be observations from a normal distribution with mean zero and variance σ_e^2.

If readers find this measurement study model confusing, it is sufficient simply to realize that the model involves three sources of variation – subject (denoted by σ_b^2), observer (denoted by σ_c^2) and residual (denoted by σ_e^2).

While the ANOVA summary displayed in table 23.2 has the same form as those we encountered in chapters 10 and 15, it is used in a different way for reliability and agreement studies. We will not discuss the different uses in general but simply say that the entries in the column labelled Mean square are used to estimate variances such as σ_b^2, σ_c^2 and σ_e^2. For example, in the particular case of table 23.2, it turns out that the mean square for subjects is an estimate of $J\sigma_b^2 + \sigma_e^2$, the mean square for observers is an estimate of $N\sigma_c^2 + \sigma_e^2$, and the mean square for residuals is an estimate of σ_e^2. By suitably manipulating these expressions, we obtain estimates of σ_b^2, σ_c^2 and σ_e^2 that are equal to 20.68, 2.75 and 15.40, respectively.

The total variation in the observations carried out in the study is the sum of the three variances, i.e., $\sigma_b^2 + \sigma_c^2 + \sigma_e^2$. If the measurement under investigation is to have any value for discriminating between subjects, then the variation associated with the subject effects should represent the largest fraction of this total variation. An intraclass correlation coefficient (ICC), which is defined as the ratio

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2 + \sigma_e^2},$$

measures this fraction. Thus, the ICC compares the component of variation in BILAG associated with subjects to the total variance in BILAG.

As we indicated earlier, the full details of how to estimate an ICC are beyond the scope of this book. Note, also, that there is no particular interest in testing the hypothesis that the ICC is equal to any particular value. However, from an appropriate statistical analysis, an estimate and confidence interval for an ICC can be calculated. These quantities can be used in the usual way to quantify what can be said about reliability based on the available data. For the example summarized in table 23.1, the estimated ICC is 20.68/(20.68 + 2.75 + 15.40), which equals 0.53. With only 53% of the total variation in BILAG scores associated with the subjects, the reliability of the activity score measurement appears to be moderate at best. The 95% confidence interval associated with this estimate is (0.28, 0.84), and the rather considerable width of this interval reflects the relatively small sample size. Therefore the true value of the ICC is rather uncertain. The reliability of the BILAG measurement could be as low as 0.28, which is quite small; on the other hand, it might also be as large as 0.84, which would indicate quite good reliability.

There are other experimental designs that one can use to undertake a measurement reliability study, and the definition of the intraclass correlation coefficient varies somewhat with different designs. We have chosen to limit our discussion to the case we described because, although it is not the simplest statistically, this design with multiple observers is a very common one that amply demonstrates the general concept that an ICC conveys.

Finally, note that the ICC is primarily an evaluation of measurement reliability, that is of the ability of a measurement to discriminate between subjects or patients. It does not explicitly address the question of agreement between measurements taken on the same subject by different observers.

23.3. Assessing Agreement

For continuous measurements, the use of ANOVA to study reliability and agreement is the basis of 'generalizability' studies [82]. In the corresponding literature, it is recommended that investigators undertake and present a full analysis of variance, including the estimation of multiple sources of variance. The square roots of the estimated variance components, which have the same dimension as the original measurements, estimate standard deviations that can be compared to evaluate the relative size of the different sources of variation on a meaningful scale.

While careful examination of a full ANOVA table is generally recommended, specific measures of reliability and agreement are designed to extract the most relevant information from the ANOVA for the measurement study. For example, reliability coefficients are often ratios of variance estimates. Since these quantities are widely known and convey useful information, they can be used as convenient summary measures of reliability when one is examining a large number of outcomes. The gain from attempting to develop an alternative measure based on standard deviations would likely be minimal.

In contrast to the question of measurement reliability, there is more debate as to what constitutes a suitable summary of measurement agreement, and there is, perhaps, a need to consider quantities that are clearly distinguishable from measures of reliability. Following Isenberg et al. [83], we suggest using the ratio of the standard deviation of measurement attributable to the observers and the standard deviation of measurement attributable to the subjects as a measure of agreement. For the measurement model for BILAG scores that we used in the previous section, this ratio of standard deviations is equal to $\sigma_c/\sigma_b = r$. A small value of r is associated both with a small value of σ_c, which indicates a high level of agreement between the observers, and a large value of σ_b. Since this summary indicator of agreement does not involve σ_e, it provides a way of assessing the level of observer/rater concurrence irrespective of the magnitude of the residual variation.

For the example we considered in §23.2, the estimates of σ_c and σ_b are 1.66 and 4.55, respectively. This gives an estimated value for r of 0.36. The corresponding 95% confidence interval for r that can be calculated from this estimate is (0, 0.97). While the small values in this confidence interval suggest that measurement agreement might be good, the larger values of r that are consistent with the study data indicate that the standard deviation for variation due to observers could have a magnitude comparable to the standard deviation associated with the patients, i.e., a value of r that is approximately 1. A larger study would be required to provide more precise information.

Table 23.3. Assessments of musculoskeletal disease activity for 80 subjects provided by two physicians

Physician 1	Physician 2		Total
	significant activity	little or no activity	
Significant activity	9	6	15
Little or no activity	7	58	65
Total	16	64	80

Table 23.4. The table of expected counts corresponding to the observed data summarized in table 23.3

Physician 1	Physician 2		Total
	significant activity	little or no activity	
Significant activity	$\frac{15\times16}{80}=3$	$\frac{15\times64}{80}=12$	15
Little or no activity	$\frac{65\times16}{80}=13$	$\frac{65\times64}{80}=52$	65
Total	16	64	80

23.4. The Kappa Coefficient

For data that are not continuous measurements, the statistical method most commonly associated with reliability and agreement is called κ, the kappa coefficient. It is usually motivated in the following fashion.

Consider first the situation where two observers each assess the same study subjects using a common binary classification scheme. Table 23.3 summarizes the assessments of musculoskeletal disease activity for 80 subjects made by two physicians.

From table 23.3 we can see that the two physicians agree on 67/80 = 83.8% of their assessments. However, even if each physician randomly chose a category for each subject, some minimal level of agreement would be observed. Therefore, κ is designed to compare observed agreement with what could be expected purely by chance.

Table 23.5. The 4 × 4 table summarizing two physicians' assessments of disease activity in 80 lupus patients

Physician 1		Physician 2				Total
		disease activity				
		A	B	C	D	
Disease activity	A	2	0	1	0	3
	B	2	5	4	1	12
	C	0	5	14	16	35
	D	0	2	3	25	30
	Total	4	12	22	42	80

Calculating the required expected values for the entries in table 23.3 is no different than the calculations we first outlined in chapter 4 for the 2 × 2 table. This is because we are assuming that the physicians' assessments are mutually independent, and therefore agreement arises by chance. The corresponding expected values are displayed in table 23.4.

In the current situation, we are only interested in the diagonal entries in the table, since these represent subjects on whom both physicians' assessments agree. For example, Physician 1 classified 15/80 (18.8%) of the patients as having active disease; for Physician 2 the corresponding result was 16/80 (20%). Therefore, the probability that both physicians would classify a given patient as having active disease, purely by chance, would be 0.188 × 0.20 = 0.038. For a group of 80 patients, the corresponding expected value appearing in the upper left corner entry of the table would be 0.038 × 80 = 3.

From table 23.4, we see that 55 agreements in the 80 patients could be expected by chance. The observed number of agreements was 67 so that 67 – 55 = 12 'extra' agreements were observed. Kappa is equal to the ratio of the observed extra agreements to the maximum possible number of extra agreements which, in this case, would equal 80 – 55 = 25. Thus the value of κ for table 23.3 would be 12/25 = 0.48, indicating that 48% of the possible increase in agreement, in excess of that due to chance, was observed.

The definition of κ is easily extended to classification schemes with more than two categories. The data in table 23.3 came from an experiment in which lupus patients were assigned to one of four categories of disease activity. The original data from the experiment are summarized in table 23.5, using the labels A, B, C, and D for the disease activity categories; A represents the highest level of activity.

If we use the same logic in this case that we did for table 23.3 and calculate expected values for the diagonal entries using the rule we outlined in §4.3, the required values will be 0.15, 1.80, 9.63 and 15.75, giving a total of 27.33 agreements that might be expected by chance. The total observed number of agreements is 2 + 5 + 14 + 25 = 46, and therefore the value of κ for table 23.5 is (46 – 27.33)/(80 – 27.33) = 0.35.

If we represent the observed and expected entries for classification categories 1, ..., K by $O_{11}, ..., O_{KK}$ and $e_{11}, ..., e_{KK}$, respectively, then

$$\kappa = \frac{\sum_{i=1}^{K} O_{ii} - \sum_{i=1}^{K} e_{ii}}{N - \sum_{i=1}^{K} e_{ii}},$$

where N represents the number of patients or subjects.

The value of κ lies between –1 and 1, although in some tables the lower limit is greater than –1. Negative values indicate that observed agreement is less than that expected by chance, suggesting that the observers tend to disagree rather than to agree.

The values of κ that represent good and poor agreement can only be assessed subjectively. A common guideline is to say that $\kappa > 0.75$ represents good agreement, and $\kappa < 0.4$ represents poor agreement. We do not recommend adopting any particular values as canonical, especially given some of the discussion in §§23.5–23.7. However, the guidelines we just mentioned may be useful in some contexts.

23.5. Weighted Kappa

The ordinary kappa statistic only counts those subjects to whom both raters assign the same classification, and any disagreements are treated with equal severity. However, in a summary such as table 23.4, the categories are naturally ordered and some investigators might claim that disagreements which belong to adjacent categories are less serious than those assigned to categories that are farther apart. This notion gives rise to a summary measure of agreement called a weighted kappa.

To calculate a weighted kappa, it is first necessary to assign 'agreement weights' to all cells in the K × K summary table. Exact agreement, which corresponds to observations recorded on the diagonal of the summary table, is assigned a weight, which we can denote by w_{ii}, of 1. Other cells will have weights that represent a fractional value. Thus, the entry belonging to row i and column j is assigned the agreement weight w_{ij} between 0 and 1. The weighted measure of observed agreement will then involve a sum over all cells of the table and can be represented by

$$A_O = \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} O_{ij}.$$

If we use the same set of assigned weights, the weighted value of expected agreements is equal to

$$A_e = \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} e_{ij}.$$

Then the weighted value of κ is equal to

$$\kappa_w = \frac{A_O - A_e}{N - A_e},$$

and is interpreted in the same way as the unweighted kappa. Readers who want to be sure that they understand this formula should check that the value of κ_w is the same as that discussed in §23.4 if $w_{ij} = 0$ when i is different from j.

A common choice of weights is the so-called quadratic weights, which are equal to

$$w_{ij} = 1 - \frac{(i-j)^2}{(K-1)^2}.$$

Another commonly-used set of weights is called the Cicchetti weights. Normally, any of the weights available in computer programs that calculate weighted kappa values will be adequate. If there is any concern about the effect of choosing a particular set of weight values, κ_w can be calculated with different sets of weights to see if any change in qualitative conclusions arises.

If we use the quadratic weights for the data summarized in table 23.4, $\kappa_w = 0.47$. This value is larger than $\kappa = 0.35$ that we calculated in §23.4, reflecting the fact that disagreements between adjacent categories are not treated as seriously as the disagreements concerning a particular patient that span three categories.

It turns out that κ_w, using quadratic weights, is equivalent to an intraclass correlation coefficient. Since the ICC is used to assess measurement reliability, it would seem that κ_w also has properties that are appropriate for a measure of reliability. This will become further apparent in light of our discussion of a particular feature of κ in §23.7.

23.6. Measures of Agreement for Discrete Data

The quantity r that we introduced in §23.3 can be adapted for use with discrete data, but the details are beyond the scope of this book. Alternatively, we can define agreement measures that are based on the concept of odds ratios – natural characterizations of 2 × 2 tables. But any discussion of this

Table 23.6. Details of hypothetical screening tests illustrating the dependence of κ on disease prevalence

		Disease status			κ
		diseased	disease-free	total	
Disease prevalence 50%					
Test status	+	150	50	200	
	–	50	150	200	0.50
Total		200	200	400	
Disease prevalence 30%					
Test status	+	30	10	40	
	–	90	270	360	0.27
Total		120	280	400	
Disease prevalence 25.5%					
Test status	+	3	1	4	
	–	99	297	376	0.02
Total		102	298	400	

approach would quickly move beyond the bounds of what is reasonable here. Therefore, if this subject of agreement is of particular interest, we are compelled to leave the reader to explore the topic elsewhere, or to consult a statistician if questions arise in a particular application.

In our discussion thus far, we have only considered the situation when observers have equal status. However, measures of agreement are also of interest when one or more methods of observation are being compared with a 'gold standard'. The evaluation of diagnostic tests, which we discussed in chapter 22, is a special case of this particular problem. However, we will use this scenario in the next section to illustrate a particular feature of κ of which the reader should be aware.

23.7. The Dependence of κ on Prevalence

To illustrate that κ falls most naturally into the category of a reliability measure, and is not purely a measure of agreement, consider a hypothetical screening test, T, for a particular disease. This represents a slightly different situation than having two raters or assessors of equal status, but is still one in which κ might well be used by some investigators.

As we discussed in chapter 22, relevant probabilities for this setting are the predictive value of the test for the disease states, i.e., the two post-test probabilities. We will assume that the post-test probability an individual is diseased, given that the screening test is positive, is 75%, and that the corresponding post-test probability an individual is disease-free, given that the screening test is negative, is also 75%. Table 23.6 summarizes the expected outcomes for a sample of 400 individuals who were tested using T when the disease prevalence in the population is equal to 50, 30 and 25.5%.

As the disease prevalence decreases by roughly one-half from 50 to 25.5%, table 23.6 shows that κ decreases from 0.5 to 0.02. Nevertheless, agreement between the test result and disease status which, in this situation, is probably most sensibly reflected by the post-test probabilities, is the same in all three scenarios. Likewise, the estimated odds ratio in each situation is equal to nine. Thus the change reflected in the marked decrease in κ appears to hinge on a reduced ability to discriminate between individuals as the disease prevalence decreases simply because the population becomes more homogeneous with respect to disease status. A statistic, such as κ, that depends on disease prevalence is a sensible choice for a measure of reliability but not for a measure of agreement.

We note that κ, and various alternatives to κ, are the continuing subject of considerable discussion. Many different views have been suggested, and some authors are quite passionate about the preferred choice. While we do have opinions on these matters, many of the pertinent issues involve concepts that are beyond the scope of this book. Nonetheless, we believe readers ought to be aware that κ depends on prevalence, and therefore exercise due caution if κ is portrayed elsewhere solely as a measure of agreement.

References

1 Thomas PRM, Tefft M, D'Angio GJ, Norkool P, Farewell VT: Relapse patterns in irradiated Second National Wilms' Tumor Study (NWTS-2) patients. Proc Am Soc Clin Oncol 1983;24:69.

2 Backhouse CI, Matthews JA: Single-dose enoxacin compared with 3-day treatment for urinary tract infection. Antimicrob Agents Chemother 1989;33:877–880.

3 Storb R, Prentice RL, Thomas ED: Marrow transplantation for treatment of aplastic anemia. N Engl J Med 1977;296:61–66.

4 Medical Research Council: Treatment of pulmonary tuberculosis with streptomycin and para-amino-salicylic acid. BMJ 1950;ii:1073–1085.

5 McDonald GB, Sharma P, Matthews D, Shulman H, Thomas ED: Venocclusive disease of the liver after bone marrow transplantation: Diagnosis, incidence and pre-disposing factors. Hepatology 1984;4:116–122.

6 Bishop YMM: Full contingency tables, logits, and split contingency tables. Biometrics 1969;25: 383–399.

7 Berkson J, Gage RP: Calculation of survival rates for cancer. Proc Staff Meet Mayo Clin 1950;25: 270–286.

8 Cutler SJ, Ederer F: Maximum utilization of the life table method in analyzing survival. J Chronic Dis 1958;8:699–712.

9 Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–481.

10 Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 1977;35:1–39.

11 James RC, Matthews DE: The donation cycle: a framework for the measurement and analysis of blood donor return behaviour. Vox Sang 1993;64:37–42.

12 James RC, Matthews DE: Analysis of blood donor return behaviour using survival regression methods. Transfus Med 1996;6:21–30.

13 Follmann DA, Goldberg MS, May L: Personal characteristics, unemployment insurance, and the duration of unemployment. J Econom 1990;45:351–366.

14 Prentice RL, Marek P: A qualitative discrepancy between censored data rank tests. Biometrics 1979;35:861–867.

15 Gehan EA: A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 1965;52:203–223.

16 Peto R, Peto J: Asymptotically efficient rank invariant test procedures (with discussion). J R Stat Soc A 1972;135:185–206.

17 Pearson K, Lee A: On the laws of inheritance in man. I. Inheritance of physical characters. Biometrika 1903;2:357–462.

18 Wainwright P, Pelkman C, Wahlsten D: The quantitative relationship between nutritional effects on preweaning growth and behavioral development in mice. Dev Psychobiol 1989;22:183–195.

19 Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research, ed 4. Oxford, Blackwell Science, 2002.

20 Duncan BBA, Zaimi F, Newman GB, Jenkins JG, Aveling W: Effect of premedication on the induction dose of thiopentone in children. Anaesthesia 1984;39:426–428.

21 Von Bortkiewicz L: Das Gesetz der kleinen Zahlen. Leipzig, Teubner, 1898.

22 Fallowfield L, Jenkins V, Farewell V, Saul J, Duffy A, Eves R: Efficacy of a Cancer Research UK communication skills training model for oncologists: a randomised controlled trial. Lancet 2002;359:650–656.

23 Breslow NE, Day NE: Statistical Methods in Cancer Research, vol II: The Design and Analysis of Cohort Studies. Lyon, International Agency for Research in Cancer, 1987.

24 Trinchieri G, Kobayashi M, Rosen M, Loudon R, Murphy M, Perussia B: Tumor necrosis factor and lymphotoxin induce differentiation of human myeloid cell lines in synergy with immune interferon. J Exp Med 1986;164:1206–1225.

25 Cox DR: Regression models and life-tables (with discussion). J R Stat Soc B 1972;34:187–220.

26 Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE: The analysis of failure times in the presence of competing risks. Biometrics 1978;34:541–554.

27 Calzavara LM, Coates RA, Raboud JM, Farewell VT, Read SE, Shepherd FA, Fanning MM, MacFadden D: Ongoing high-risk sexual behaviors in relation to recreational drug use in sexual encounters. Ann Epidemiol 1993;3:272–280.

28 Liang KY, Zeger S: Longitudinal data analysis using generalized linear models. Biometrika 1986; 73:13–22.

29 Zeger SL, Liang KY: Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–130.

30 Gladman DD, Farewell VT, Nadeau C: Clinical indicators of progression in psoriatic arthritis (PSA): a multivariate relative risk model. J Rheumatol 1995;22:675–679.

31 Kalbfleisch JD, Lawless JF: The analysis of panel data under a Markov assumption. J Am Stat Assoc 1985;80:863–871.

32 Mantle MJ, Greenwood RM, Currey HLF: Backache in pregnancy. Rheumatol Rehabil 1977;16: 95–101.

33 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.

34 Freedman LS: Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1982;1:121–129.

35 Pocock SJ: Clinical Trials, a Practical Approach. Chichester, Wiley, 1983.

36 Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585–612.

37 Breslow N: Perspectives on the statistician's role in cooperative clinical research. Cancer 1978; 41:326–332.

38 Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, Gail MH, Ware JH: Randomized clinical trials. N Engl J Med 1976;295:74–80.

39 Chalmers TC: Randomized clinical trials in surgery; in Varco RL, Delaney JP (eds): Controversy in Surgery. Philadelphia, Saunders, 1976, pp 3–11.

40 Chalmers TC, Block JB, Lee S: Controlled studies in clinical cancer research. N Engl J Med 1972; 287:75–78.

41 Cox DR: Summary views: a statistician's perspective. Cancer Treat Rep 1980;64:533–535.

42 Farewell VT, D'Angio GJ: A simulated study of historical controls using real data. Biometrics 1981;37:169–176.

43 Freireich EJ: The randomized clinical trial as an obstacle to clinical research; in Delaney JP, Varco RL (eds): Controversies in Surgery. II. Philadelphia, Saunders, 1983, pp 5–12.

44 Gehan EA, Freireich EJ: Non-randomized controls in cancer clinical trials. N Engl J Med 1974; 290:198–203.

45 Sacks H, Chalmers TC, Smith H: Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–240.

46 Pocock SJ, Elbourne DR: Randomized trials or observational tribulations. N Engl J Med 2000; 342:1907–1909.

47 Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG: Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii–x, 1–173.

48 Yule GU: The Function of Scientific Method in Scientific Investigation. Industrial Fatigue Research Board Report 28. London, HMSO, 1924.

49 Irwin JO: The place of mathematics in medical and biological statistics. J R Stat Soc A 1963;126: 1–45.

50 McPherson K: Statistics: the problem of examining accumulating data more than once. N Engl J Med 1974;290:501–502.

51 Whitehead J: The Design and Analysis of Sequential Clinical Trials, ed 2. New York, Horwood, 1991.

52 Fleming TR, Harrrington DP, O'Brien PC: Designs for group sequential tests. Control Clin Trials 1984;5:348–361.

53 Prentice RL: Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431–440.

54 Fleming TR: Evaluating therapeutic interventions: some issues and experiences. Stat Sci 1992;7: 428–456.

55 Burzykowski T, Molenberghs G, Buyse M (eds): The Evaluation of Surrogate Endpoints. New York, Springer, 2005.

56 Fleming TR: Evaluation of active control trials in AIDS. J Acquir Immune Defic Syndr 1990;3: S82–S87.

57 Cox DR: A remark on multiple comparison methods. Technometrics 1965;7:223–224.

58 Pocock SJ, Geller NL, Tsiatis AA: The analusis of multiple endpoints in clinical trials. Biometrics 1987;43:487–498.

59 Torrance GW: Utility approach to measuring health-related quality of life. J Chronic Dis 1987; 40:593–600.

60 Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR: Quality of life assessment: can we keep it simple? J R Stat Soc A 1992;155:353–393.

61 Cook RJ, Farewell VT: Guidelines for monitoring efficacy and toxicity responses in clinical trials. Biometrics 1994;50:1146–1152.

62 DeMets DL: Stopping guidelines vs. stopping rules: a practitioner's point of view. Commun Stat Theory Methods 1984;13:2395–2417.

63 Cook RJ: Coupled error spending functions for parallel bivariate sequential tests. Biometrics 1996;52:442–450.

64 Peto R, Collins R, Gray R: Large-scale randomized evidence: large, simple trials and overviews of trials; in Warren KS, Mosteller F (eds): Doing More Good Than Harm: The Evaluation of Health Care Interventions. Ann NY Acad Sci 1993;703:314–340.

65 Lilienfeld AM, Lilienfeld DE: Foundations of Epidemiology, ed 3. New York, Oxford University Press, 1994.

66 Prentice RL, Shimizu Y, Lin CH, Peterson AV, Kato H, Mason MW, Szatrowski TP: Serial blood pressure measurements and cardiovascular disease in a Japanese cohort. Am J Epidemiol 1982; 116:1–28.

67 Weiss NS, Szekeley DR, English DR, Schweid AJ: Endometrial cancer in relation to patterns of menopausal estrogen use. JAMA 1979;242:261–264.

68 Breslow NE, Day NE: Statistical Methods in Cancer Research, vol 1: The Analysis of Case-Control Studies. Lyon, International Agency for Research in Cancer, 1980.

69 Hauck WW: The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics 1979;35:817–819.

70 Robins J, Breslow N, Greenland S: Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics 1986;42:311–323.

71 Doll R, Hill AB: Mortality of British doctors in relation to smoking: observations on coronary thrombosis. Natl Cancer Inst Monogr 1996;19:205–268.

72 McNeil D: Epidemiological Research Methods. New York, Wiley, 1996.

73 Weiss NS: Clinical Epidemiology: The Study of the Outcome of Illness, ed 2. New York, Oxford University Press, 1996.

74 Catalona WJ, Hudson MA, Scardino PT, Richie JP, Ahmann FR, Flanigan RC, deKernion JB, Ratliff TL, Kavoussi LR, Dalkin BL, et al: Selection of optimal prostate-specific antigen cutoffs for early detection of prostate cancer: receiver operating characteristic curves. J Urol 1994;152: 2037–2042.

75 Ransohoff DF, Feinstein AR: Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–930.

76 Reid MC, Lachs MS, Feinstein AR: Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274:645–651.

77 Lusted LB: Introduction to Medical Decision Making. Springfield, Thomas, 1968.

78 Sackett DL, Haynes RB, Guyatt GH, Tugwell P: Clinical Epidemiology: A Basic Science for Clinical Medicine, ed 2. London, Little, Brown, 1991.

79 Fagan TJ: Nomogram for Bayes' theorem. N Engl J Med 1975;293:257–261.

80 Guyatt GH, Oxman AD, Ali M, Willan A, Mcllroy W, Patterson C: Laboratory diagnosis of iron-deficiency anaemia: an overview. J Gen Intern Med 1992;7:145–153.

81 Peabody GW: The physician and the laboratory. Boston Med Surg J 1922;187:324–327.

82 De Vet H: Observer reliability and agreement; in Armitage P, Colton T (eds): Encyclopedia of Biostatistics, ed 2. Chichester, Wiley, 2005, pp 3801–3805.

83 Isenberg DA, Allen E, Farewell VT, Ehrenstein MR, Hanna MG, Lundberg IE, Oddis D, Pilkinton C, Plotz P, Scott D, Vencovsky J, Cooper R, Rider L, Miller F: International consensus outcome measures for patients with idiopathic inflammatory myopathies. Development and initial validation of myositis activity and damage indices in patients with adult onset disease. Rheumatology 2004;43:49–54.

Subject Index

Accuracy of χ^2 approximation 34–35
Actuarial curves 55, *see also* Kaplan-Meier
Adjustment
  for confounding 47–48, 52–53, 155
  for covariates 117, 125, 130, 163, 198
  for heterogeneity in 2 × 2 tables 46–51, 133–135
  for prognostic factors 69
Agreement and reliability 298–309
  observer agreement 299
  observer reliability 299
Alpha (α) 209, 210, 211, 230–231, 246–247
Analysis
  initial 193–196
  primary 196–200
  secondary 200–201
Analysis of variance 118, 124–127, 174–190
  as linear regression with categorical covariates 175–181
  components of total variation 123
  degrees of freedom 123, 179
  explained/residual variation 122–123
  F test 124–127, 179
  lack of fit 127
  main effects model 180–181
  mean square 124, 301
  pure error 126–127
  R^2 124
  repeated measures 160, 185
  repeat observations 126
    on same subject 160
  sum of squares 123, 179, 186–187
    between patients 186
    lack of fit 127
    model 123
    pure error 126–127
    residual 123
    total 123
    within patients 186
  variance estimate 126
  with observer effects 300
  with subject effects 300
  with two-factor interactions 181–183, 187
ANOVA tables 118, 122–127, 174, 179, 186–187
Association 111, 119, 130, 163

Bar graph, *see* Histogram
Beta (β) 209–211
Binary data 19
  paired 46, 51–53
Binary variable 128
  expected value 129
Binomial coefficient 21
Bonferroni adjustment 202, 206, 241

Case-control study 135, 264, 268–273
  comparison of disease rates 264
Categorical measurements 174–175, *see also* Categories
Categorical variables, *see* Categorical measurements
Categories
  grouping 196–197, 204
  response variable 196
  using indicator variables 175–181, 181–183, 196–197, 276
Caterpillar plot, *see* Meta–analysis, graphical displays
Censored survival times 56, 59, 60, 149
  independence assumption 60
Central limit theorem 81, 82
Chance 16, 17
Chi-squared (χ^2) distribution 33, 34, 40, 42–45, 69
  link with normal 83–84
  mean 42
  variance 42
Chi-squared (χ^2) test
  for 2 × 2 tables 28–35, 202
  for rectangular tables 35–42, 203
  increasing significance 204
Chronic diseases 149
Clinical trials, *see also* Design of clinical trials
  conditional power 252
  equivalence 236–237
  hazards of small 214–216
  intention to treat analysis 227
  non-inferiority 237
  sequential analysis 230–233
    for a single outcome 230–233
    for efficacy-toxicity outcomes 244–250
  size, *see* Sample size
  stochastic curtailment 250–253
  use of orthogonal contrasts 243
Cohort
  population-based 264
  study 263
Column totals 20
Combining 2 × 2 tables 46–51
Combining rows and columns 38, 40–42
Common variance assumption 103, 106–110
Comparable events, set 18, 21
Comparison of survival curves 67–75
  adjusting for other factors 69
  restricted 70
Comparisons with normal data 100–110
  paired 100–103
  unpaired 103–106
Confidence intervals 84–87
  95% 85, 86
  based on t distribution 97–100
  definition 84, 85
  difference in population means 104–106
  in equivalence trials 236–237
  information provided by 85, 86
  length 85–86
  level of confidence 85
  mean of normally distributed data 96–98
  odds ratio 132
  regression coefficients 132, 144, 153
  relative risk 153
Confounding factor 47–51, 270
Consistency
  between data and hypothesis 11, 14–15, 17
  with null hypothesis 17, 18
Contingency table
  2 × 2 12, 28–35
  rectangular 35–42
Continuity correction, *see* Yates
Controls, *see* Design of clinical trials
Correlation 118–122, 162
  coefficient 119
  effect of selection 120
  in longitudinal studies 160, 162, 164–165
  versus regression 118–122
Covariate 113, 129, 150, 168–169, *see also* Explanatory variable
  fixed 158
  time-dependent 157–159, 265
Cox regression model, *see* Proportional hazards regression
Critical values
  χ^2 distribution 42–45
  F-distribution 106, 107–108
  normal distribution 83
  t distribution 98, 99

Cumulative hazard 157
Cumulative mortality figure 145
Cumulative probability curve,
 see Probability curve
Cured fraction of patients 59–60

Data
 analysis 191–206
 cleaning up 193–196
 missing 192
 quality 191–192
 retrospective/prospective collection
  192–193, 263–264
Death rate, *see also* Hazard rate
 at time t 150
 regression model 150
 relation to survival curve 151
Degrees of freedom, *see also* Analysis of
 variance
 χ^2 distribution 42
 F distribution 106
 t distribution 97–98
Delta (δ) 104, 210, 236
 clinically relevant difference 210
 departure from therapeutic equivalence
  236–237
 difference in means 104
Design of clinical trials 217–253
 active control 236–237
 cluster randomized study 238
 cross-over trial 238
 data collection 218
 Data Monitoring Committee 222
 early stopping, *see* sequential designs
 efficacy-toxicity 244–250
 equivalence, *see* active control
 ethics committee 221
 factorial designs 228
 group randomized trial 238
 group sequential designs 231–233,
  244–250
 multiple outcomes 239–242, 244
  summary response 240, 242
 multiple treatment arms 242–244
 n-of-1 trial 238
 parallel group design 237
 questions to be answered 217
 randomized versus historical controls
  222–227
 sequential designs 230–233, 244–250
 treatment assignment 218, 219
 treatment selection 218
 Trial Management Committee 221
 trial organization 220–222
 trial protocol 221
 Trial Steering Committee 221
Deviance 205
Diagnostic test 282–297
 false negative error 287
 false negative outcome 286
 false positive error 287
 false positive outcome 286
 likelihood ratios 292
  of a negative test result 293
  of a positive test result 292
 post-test probability of disease 288
  converting prevalence to 288–292,
   293–295
 post-test probability of no disease 288
  converting prevalence to 288–292,
   294–295
 predictive value of a negative test
  outcome 288
 predictive value of a positive test
  outcome 288
 pretest probability of disease 288
 prevalence 288
 sensitivity 286–287
 specificity 287
 spectrum of study subjects 287–288
 true negative outcome 286
 true positive outcome 286
 use of binary outcomes 284
Discordant pairs 53
Disease incidence
 cohort data 263
 rate of 264
 regression equation 264, 265
Dispersion 10, *see also* Variance
Distribution, *see* Probability distribution
 explanatory variables 116
 spread 10
 survival times 149, 151
Dose-response study 137, *see also* Logistic
 regression, Poisson regression

ED_{50}, *see* Median effective dose
Effect graph 188, *see also* Interaction, graphical interpretation of
Effect modifying factor 270, 271, 278–280, *see also* Interaction
Epidemiological studies 263–264, *see also* Cohort, Case-control
Epidemiology 135, 263, 280–281
  clinical 280–281
Estimate
  approximate asymptotic distribution 82
  notation 82, 113
  precision 86
Estimated standard deviation 82, 95
Estimated standard error 82
Estimation 54, 84
  in 2 × 2 tables 131–135
Examples
  abnormal ovulatory cycles and diabetes 164–167
  anti-tumor activity in leukemic mice 23–25
  assessments of musculoskeletal disease by two physicians 304–306, 306–307
  autologous marrow transplants in dogs 25–26
  back pain in pregnancy 193–201, 202–204
  BILAG scores for systemic lupus erythematosus 300-302, 303
  Cardiac Arrhythmia Suppression Trial 235
  case-control study of endometrial cancer 268–270, 271–273
  cell differentiation and immunoactivation by TNF and IFN 145–148
  cohort study of cardiovascular disease 265–268
  cohort study of high-risk sexual activity 161–164
  cohort study of smoking and coronary deaths 275–280
  comparing the return behaviours of whole blood donors 75
  coronary artery disease 225–226
  diabetes and HLA 202, 204, 205
  diagnostic consistency between pathologists 51–53
  dose requirement of anesthetic 136–140
  efficacy-toxicity study in kidney transplant patients 249–250
  γ-interferon in chronic granulomatous disease 235
  heights of father-son pairs 112–113
  immunological assay 90–109
    effect of concentration 100–103
    hemophiliacs versus controls 104–109
  joint damage in psoriatic arthritis 167–173
  leukemic relapse and GVHD 158–159
  marrow transplants for aplastic anemia 29–34, 128–129, 130, 131–135
  nutritional effects on mouse pups 113–118, 119–120, 122–127
  paired enzyme-linked immunosorbent assays (ELISAs) 282–284
  PAS and streptomycin for tuberculosis 35–40
  physician communication skills 141–142, 144
  prenatal care 47–51, 135–136
  prothrombin time international normalized ratio (INR) 174–190
  randomized clinical trials of prolonged antiplatelet therapy 257–262
  return behaviour of whole blood donors 64–66, 75
  serum ferritin concentration for iron deficiency anemia 295–296
  small-cell lung cancer 233
  survival for lymphoma patients 54–55, 61, 67–68, 74, 81–82, 83, 86, 150, 152–157
  treating urinary tract infection with enoxacin 26–27
  venocclusive disease of liver 40–42
  Wilms' tumor
    chemotherapy 223–225, 228
    inadequate radiotherapy/relapse 12–17
    study design 228

Expected numbers
calculating κ 305, 306
in 2 × 2 tables 30–31
in rectangular tables 37, 38
of deaths in log-rank test 68–69, 71–73
too small 38, 41
Explanations for observed result 16
Explanatory variable 113, 194, *see also* Covariate, Risk factor
additional information 117, 125
identification of important 196–200
joint effects 117, 124–125
Exploratory analysis 193–196
Exposure variable, *see* Risk factor
alternative coding for 272–273

Factor 67, 69, 100, 174, 181–183
Failure 19
False discovery rate 206
False negative 209, *see also* Type II error, Diagnostic test
and surrogate endpoints 235
False positive 209, 240, 247, *see also* Type I error, Diagnostic test
and surrogate endpoints 235
F distribution 106–110, 124, 179
Fisher's test 19–27
approximate 28–35
assumptions 20, 29
null hypothesis 21
Follow-up 218
Force of mortality 150, *see also* Hazard rate
Frequency 3
Futility index 252

Galton 112
Generalizability studies 303
Generalized estimating equations (GEE) 162, *see also* Regression model
Global test 124–125, 179, 202, 205, 206, 242
Graphs 194

Hazard rate 150
Health economic measures 242
Heterogeneity 111
Histogram 3–5, 93-94, 101–102
Hypothesis, *see* Null
generation 201

Incomplete observation 149
Indicator variables, *see* Categories
Intention to treat, *see* Clinical trials
Interaction
and effect modification 271, 278–280
as a product of indicator variables 183
as synergistic action in cell differentiation 146
between covariates 139, 153, 271–273
effect graph for 188
graphical interpretation of 188
two-factor 181–183
Interval scale 3
Intraclass correlation coefficient (ICC) 300-302
definition 302
estimate 302

Kaplan–Meier survival curve 54–66, 67–68, 149, 251
computation 62–64
dropping to zero 58
estimating median survival 59
general features 56–61
multiplication of probabilities 58
staircase appearance 58
standard errors 60
undefined 58, 59
Kappa (κ) coefficient 304–306
dependence on prevalence 308–309
for multiple measurement categories 305–306
for two measurement categories 304–305
weighted κ 306–307
Cicchetti weights 307
quadratic weights 307
relation to ICC 307

Lambda (λ) 142, 143, 276
Poisson mean 142, 143, 276
regression model for log λ 143, 276
Likelihood ratio statistic 205, 259, 272, 278
Likelihood ratio test 259, 278

Location, measure 8–10
Log-likelihood 205, 272, 273
Log-rank test 67–75, 149, 212, 233
  details 68–71
  differences designed to detect 67, 70
  effectiveness 70
  for k groups 71
  null hypothesis 67–68
  restricted time interval 70
  simple example 71–74
  stratified 69
  test statistic 69
Logistic regression 128–140, 164–167, 196–200, 200–201, 204–205, 257–260, 268–273
  conditional analysis 270
  for 2 × 2 tables 131–136
  for abnormal ovulatory cycles 164–167
  for case-control data 268–273
  for dose-response data 136–140
  for meta-analysis of binary outcome data 257–260
  model 129–130
  summary precautions 199–200
  with random effects 165, 259
  with stratification 133–135, 258, 269
Logistic transform of probability 129
Longitudinal data 160–173, *see also* Correlation

Mantel-Haenszel χ^2 statistic 273–275
Mantel-Haenszel estimate 274
  variance 274
Mantel-Haenszel test, *see* Log-rank test
Maximum likelihood estimation 130, 144
Mean 8–10
  estimation 94
  population 94, 95
  sample 94, 96, 103
  value 9
Measurement agreement 303
  estimate 303
Measurement repeatability, reproducibility 284
Median 9, 10
  sample median survival time 59
Median effective dose 139
  confidence interval 139
  estimation 139
Meta-analysis 254–262
  aim 255
  fixed effects versus random effects 259–260
  graphical displays 260–262
  sensitivity to assumptions 262
  study heterogeneity 256–257
    qualitative versus quantitative 257
  types of data 255–256
Minimization
  and randomization in clinical trials 219
  as protection against selection bias 219
Model, statistical 111–112
Mu (μ) 9, 10, 94, 95, 96, 97, 98, 101, 103, 104, 105, 109, 231, 247
Multi-factor 111
  statistical analysis 1
Multiple comparisons 199, 202–205, 220, 241
Multi-state regression model 167–173
  stratified 171–173

Non-randomized studies
  possible consequences of 226
Normal data 76, 90–110
  checking assumption 91–94, 101–102
  comparing means 103–106
    independent samples 103–106
    with unequal variances 109–110
Normal deviate 83
Normal distribution 5, 76–110
  and confidence intervals 84–87
  and significance testing 81–84
  bell-shaped curve 77
  calculation of probabilities 77, 80, 87–89
  cumulative probability curve 5, 77, 87
  importance 81
  mean 77
  notation 77
  of test statistics 81, 82
  probability curve 77–80, 94, 101
  standardized 77, 80, 83, 87–89
  standardizing transformation 80, 82, 89, 96

use in advanced methodology 81–82, 84–87
variance 77
Normal tables, use 87–89
Normal test statistic 82
critical values 83
Null hypothesis 17, 18, 67, 208, 209
χ^2 test in a rectangular contingency table 36
comparing means 104–105
Fisher's test 20–21, 23, 29, 36
log-rank test 67–68
matched pairs binary data 53
paired normal data 101, 103
rejecting 208, 231
when combining 2 × 2 tables 49

Odds ratio 132, *see also* Case-control study
and relative risk 135
common, *see* Mantel-Haenszel estimate
estimation 132, 135
in a dose-response study 140
in stratified 2 × 2 tables 133, 134
model 268–270
regression equation 268, 269
test for constant 134–135
Ordered variable 196
Orthogonal contrasts, *see* Clinical trials
Outcome variable 111, 113, 118, 128, 160, 168 *see also* Response variable

Paired data
binary 46, 51–53
normal 100–103
Panel data 170
Parameter 85
Person-years of follow-up 276
Phi
in active control trials, *see* (ϕ)
($\Phi(z)$) 87, 88
(ϕ) 236–237
relative efficacy, *see* (ϕ)
Poisson distribution 141
deaths due to horsekicks 141
mean 142
Poisson regression 141–148
dose response study 145–148, *see also* Examples, cell differentiation and immunoactivation by TNF and IFN
evaluating model fit graphically 148
for cohort studies 145
for death rate 276
for log λ 143
interpreted as relative risk 277
rate ratio 144
relation to logistic regression 145
Population 2, 3, 5, 12, 19, 29, 35, 54, 67, 70, 81, 95, 96, 100, 106, 120
mean 94, 96, 105, *see also* Mu
variance 94, 95, *see also* Variance
versus sample 12–13
Power 211, *see also* Clinical trials
conditional 252
Probability
= area equation 5
calculation 5, 7–8
calculation for 2 × 2 tables 21
right-hand tail 42
success and failure 19–20
Probability curve 5
cumulative 5, 7
Probability density function 5, *see also* Probability curve
Probability distribution 2–8, *see also* Statistical tables
χ^2 33, 42–45
F 106, 107–108, 124
normal 5, 76–89, 90–110
Poisson 141, 142
t 97–100
test statistic 18, 33
Proportion 19, 145, 211
Proportional hazards 154
testing 157
Proportional hazards regression 149–159, *see also* Regression model
example 152–157, 158–159, 264–268
stratified 155–157, 265
Psi (ψ) 274, 275
p-value 1, 16, 17, 18
Bonferroni adjustment 202, 241
most significant 202
precise 87

- Quality-of-life 240, 242
  - measures 242
  - studies 240

- Random effects models 164–167, 237–239
  - estimated mean 166–167
  - for clustering 239
  - for clustering by health professional 239
  - for meta-analysis 259–260
  - subject-specific parameters 165
- Random sample 3
- Random variable 2–5
  - continuous 3, 5
  - discrete 3
- Ranking of 2 × 2 tables 14
- Rate 19
- Regression 112–113
  - coefficient 113, 131, 143, 151, 153, 162, 163, 169, 170
    - equal to 0, *see* Response variable
  - equation 113
  - linear 113
  - Y on X versus X on Y 121
- Regression model
  - Cox 151
  - for count data 143, *see also* Regression model, Poisson
  - for log λ 143
  - for survival data 149
  - GEE 162
  - Liang-Zeger 161–164
  - linear 111–127, 149
  - logistic 128–140
  - main effects 180
    - simplifying a more complex model 182
  - marginal 164, 165
  - Poisson 141–148, 275–280
  - population-averaged 164
  - proportional hazards 149–159
  - simple linear 114
  - subject-specific 165
  - testing for additivity 153
  - validity 154
  - with categorical covariates 174–181
  - with subject-specific random effects 165
  - with two-factor interactions 181–183
- Relapse or death, time 149
- Relapse rate
  - leukemia 158
  - population 12–16
  - regression equation 158–159
  - sample 13
- Relative efficacy 236, 245
  - experimental treatment to active control 236, 245
  - placebo control to standard treatment 236
- Relative risk 135, 152–153, *see also* Poisson regression
  - confidence interval 153
  - in multi-state models 170
  - models 264–268
- Relative toxicity
  - experimental treatment to control therapy 245
- Repeated measurement 160, *see also* Longitudinal data
- Response variable 113, 114, 118, 143, 194
  - regression coefficient equal to 0 116
- Risk factor 263
  - relation to disease incidence 263
- Robustness 91
- Row totals 20
  - fixed and random 34

- s^2, variance estimate 95
  - pooled estimate 104
- Sample 3, 13
- Sample size
  - calculations 207–216
    - examples 211–214
    - general aspects 207–211
  - inadequate 207
  - using software packages 211
- Sampling uncertainty 16, 17, 18
- Sequential analysis 230–233, 244–250
  - stopping criterion 231
- Sigma
  - (Σ) 3
  - (σ) 10
- Significance 1
- Significance level 12, 16, 17, 18, *see also* p-value

discounting 201
exact 28
in efficacy-toxicity trials 247
small 16
Significance tests 11–18
common features 17–18
one-sided 245–246
repeated 228–230
Standard deviation 10, *see also* Standard error
sample 95
Standard error 60, 82
Standardized mortality ratio 145
Statistical calculations, formalism 191, 200
Statistical tables 8, 21
χ^2 distribution 44
F distribution 107–108
normal distribution 83, 88
t distribution 99
Stochastic curtailment, *see* Clinical trials
Stratification 46, 48, 69, 100, 111, 133
treatment assignment 218–219
Subscripted letters 2
Success 19
factors influencing probability 47
Sum of squares, *see* Analysis of variance
Summary response, *see* Design of clinical trials
Surrogate endpoints/markers 234–236
definition of 235
hazards of using 235
Survival curve 54, 151, *see also* Kaplan-Meier
Survival data, *see* Kaplan-Meier, Proportional hazards regression
characteristics 149
Survival function 54, 151
Systematic review, *see* Meta-analysis

t distribution 97–99
t test, *see* Comparisons with normal data
paired 100–103
single-sample 97–100
two-sample 103–106
Tables, *see* Contingency, Statistical tables
Tabulation 193, 194, 195
Test of significance 11–18, *see also* Significance tests
Test statistic 18
χ^2 test in 2 × 2 table 31–32
χ^2 test in rectangular table 39
comparing log–likelihoods 205
comparing means 104
comparing variances 106
F test in ANOVA tables 124, 125, 126
Fisher's test 21
paired binary data 52–53
paired normal data 101
testing mean of normal data 98
testing regression coefficients equal to zero 116, 130, 143
Transforming data 94, 101, 161
Treatment comparison 196, 226
overestimates of difference 225
Type I error 209, 240
false positive 209, 240
in selection procedures 244
Type II error 209
false negative 209

Variable, *see* Random
dependent 113, *see also* Response variable
explanatory 113, 114, 116, 117, 143, 146, 194, *see also* Covariate
independent 113
outcome, *see* Outcome variable
Variance 10
estimation 95, 104
testing for equality 106–110

Wilcoxon test 70
Within patient comparisons 187

$\bar{x}$, sample mean 94, 95

Yates' continuity correction 32
Youden 81

Z statistic 84